DocBlocks: Communication-minded Visualization of Topics in U.S. Congressional Bills

Yannick Assogba
IBM Research
1 Rogers St
Cambridge, MA 02141 USA
yannick@us.ibm.com

Irene Ros
IBM Research
1 Rogers St
Cambridge, MA 02141 USA
iros@us.ibm.com

Matt McKeon
IBM Research
1 Rogers St
Cambridge, MA 02141 USA
matt.mckeon@us.ibm.com

Copyright is held by the author/owner(s).

CHI 2010, April 10–15, 2010, Atlanta, Georgia, USA.
ACM 978-1-60558-930-5/10/04.

Abstract

US Federal legislation is a hot topic for discussion and advocacy on the web. Yet legislative bills present a significant challenge for both experts and average citizens to navigate and understand. To explore solutions to this problem, we have created DocBlocks: a prototype visualization and website that enables users to explore the content of congressional bills and communicate their findings to others. Our technique enables us to take any document from a categorized corpus, classify its sections, and visualize its topic structure. With the launch of this service, we hope to provide a valuable tool for open governance and learn from our users at this critical intersection of visualization, advocacy, social software, and civil society.

Keywords

Information visualization, government, web, legislation, social software, text classification

ACM Classification Keywords

H5.2. User Interfaces, H5.3 Group and Organization Interfaces

General Terms

Human Factors

Introduction

The day after his inauguration in January of 2009, US president Barack Obama declared that transparency and accountability would be a hallmark of his administration. In the wake of this announcement, a number of organizations have turned to the Web as a means of promoting open analysis and insight into the operation of the US government. Projects such as GovTrack.us [1], MAPLight [3], and the Sunlight Foundation [4] have emerged to act as clearinghouses for information and advocate for an ever more data-driven approaches to US politics.

Navigating the complexity of US Federal legislation is one of the many issues these organizations seek to address. A single bill may cover a wide variety of topics – from medical care to unemployment to consumer credit. Sometimes, a bill may contain elements that are unrelated to its overall subject. One of our favorite examples appears in HR627: The Credit CARD Act of 2009 [2], which imposes transparency and disclosure requirements on credit card companies. However, embedded in "Title V – Miscellaneous Provisions" is Section 512, "Protecting Americans From Violent Crime", which guarantees the right of citizens to carry guns in National Parks and Wildlife Refuges.

We have taken up the challenge of helping individuals read, explore and discuss US Federal legislation. Our prototype, *DocBlocks*, is a visualization and website that enables anyone to search and explore the content of most of the bills brought before the 111th Congress of the United States. We use an analytical method that derive fine-grained topic overviews of documents from any categorized corpus, and we then visualize the results at various levels of detail including full-text

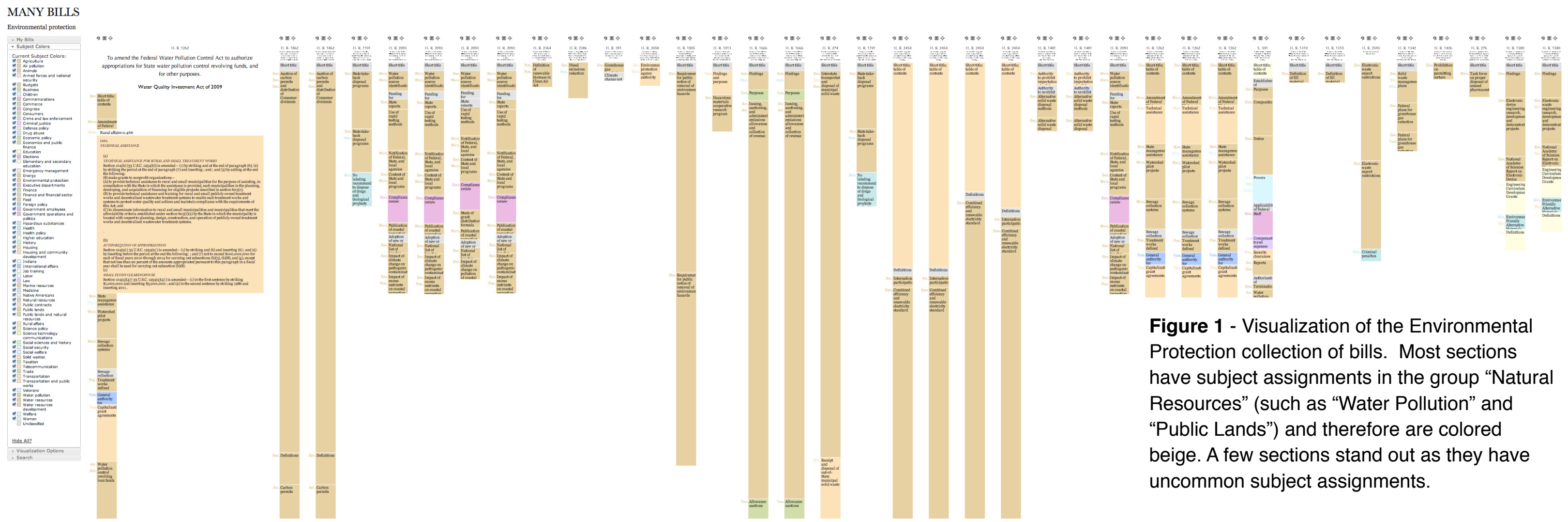
MANY BILLS
Environmental protection
To amend the Federal Water Pollution Control Act to authorize appropriations for State water pollution control revolving funds, and for other purposes.
Water Quality Investment Act of 2009

Figure 1 - Visualization of the Environmental Protection collection of bills. Most sections have subject assignments in the group "Natural Resources" (such as "Water Pollution" and "Public Lands") and therefore are colored beige. A few sections stand out as they have uncommon subject assignments.

browsing. Users can easily create custom views of collections of bills and embed them into a variety of other online media (such as blogs or social websites) for further discussion.

Background

Our own work on Many Eyes [11] and Many Eyes Wikified [10] showed that easy to use, communication-minded data visualization can serve as a valuable community component [6] for discovery, advocacy, and discussion.

Previous work on overview visualizations of large bodies of documents includes Wise et al's Galaxies and Themescapes [12], which were some of the first document visualizations to permit thematic overviews of groups of documents. Our work visualizes the thematic *content* of documents within large document corpora, as well as providing an interface to read the documents themselves.

Hearst's TileBars [8] is an early visual interface into the topical substructure of documents. While it does not identify what topics are contained in document segments, it does use colored tile sequences to illustrate term frequency in those segments for a given search query. Seesoft by Eick, Steffen & Sumner [7] is a tool that visualizes line oriented statistics related to software development, such as the date that a line of code was added or the number of times it has been changed, providing a visual summary of these statistics for several code files simultaneously. Boguraev et al's ViewTool [5] combines a high-level, topic-based document overview, and full text view into a single interface. Our visualization aims to combine a view of large numbers of documents, such as those found in TileBars and Seesoft, with the display of the thematic substructure of documents suggested by tools such as Boguraev's.

Analysis

Our text analysis focused on exposing the topical substructure of bills. As presented to Congress, a bill is formatted in discrete units called "Sections", each of which typically pertains to a single aspect of law. Therefore we chose the Section as our unit of analysis. We use a trained classifier to estimate the probability of a section of a bill being about a particular topic. To train the classifier we take advantage of the fact that each bill is assigned a "top subject" by the Congressional Research Service (CRS) of the Library of Congress. We used these subject assignments on a collection of over 40,000 bills from the past two years to train a classifier that we then use to classify individual sections. We use McAllum's MALLET toolkit [9] to perform the actual classification. Our assumption is that the model that predicts the top subject of an entire bill can be used to predict the subject of an individual section. We have not formally evaluated the efficacy of this assumption, but initial observation is very promising. For example in the aforementioned HR627 bill (the credit CARD act of 2009) while most sections carry the label "Finance and Financial Section" the section on gun holding rights is classified as "Criminal Justice" and stands out in the visualization.

Each section receives a probability for each top subject provided by the CRS. We do not attempt to classify sections with less than 50 words, as we have found that the results are rarely meaningful.

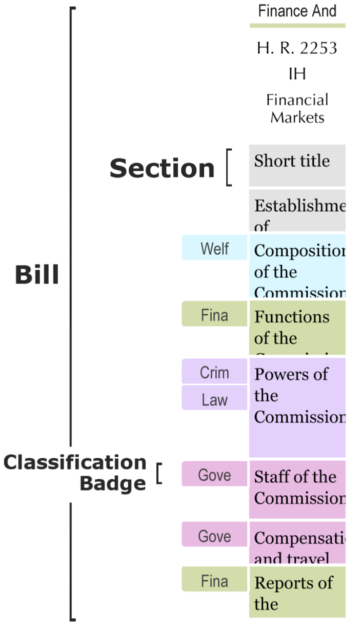

Figure 2 - A single bill as shown in our visualization.

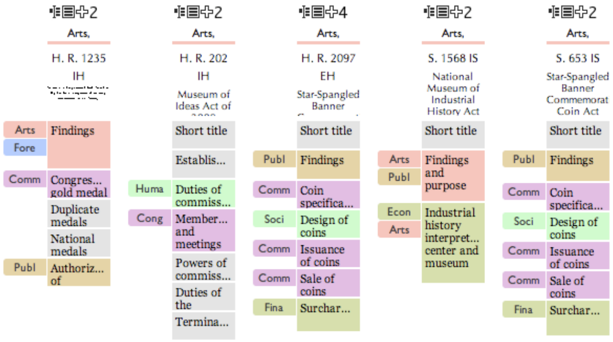

Figure 3 - A subset of bills from the Arts, Culture and Religion Collection

Design

Visualization

Our analysis generates a set of congressional bills where each section has been assigned one or more subjects, along with their associated probabilities. One visualization approach might be to assign a unique color to each subject; however, as we have over 100 different possible subjects, it is impossible to create a color scheme that allows the user to visually distinguish subjects from one another. We manually group semantically related subjects together and assign each group a color; for example, the subjects "Consumers", "Trade", and "Taxation" all fall under the "Economics" group, which is assigned a green color. We then render each bill's sections as a series of blocks; a block's color corresponds to the *group* to which the *top subject* (the subject with the highest probability) assigned to that section belongs.

The actual subject assigned to the section is shown by a small badge displayed to the left of the block; each badge is formed from the first few letters of its subject. The height of the blocks is mapped to the length of that section in the bill. Inside each section, its title is displayed to provide a quick overview of the section's content (figure 2).

The DocBlocks visualization presents a 'collection' of bills, in which each bill is displayed as a column (figure 3). Currently we display about 100 bills at a time with large collections spanning over multiple pages. We have seeded the DocBlocks website with collections of bills sharing a top subject. For the most part, bills that share a top subject will contain a similar combination of topics. By presenting collections of bills in this manner, we hope to aid the user in understanding what color combinations may be common or unusual.

One of the challenges introduced by our analysis was the variability in the subject probabilities assigned by the classifier. A single section is sometimes assigned high probabilities for multiple classifications while at other times, even its top classification is weak. Currently we try to show all the classification badges (rendered to the left of each section) that 'compete' for the top spot; we do this by ordering the corresponding classifications by score, determining the largest "gap" in the score distribution, and displaying all the classification badges above that gap. In future we hope to explore additional visual methods that would represent our level of certainty more accurately.

The visualization also includes several interactive features. When a user clicks on a block it expands and displays the full text of the section. The entire contents of a bill can be downloaded by clicking a button in the toolbar displayed above each bill. Bills may also be rearranged on the horizontal axis (for example, to compare bills side-by-side) by dragging and dropping. Users may click on the legend to turn a classification "on" or "off"; this will toggle all sections in the current view between the classification color and a neutral gray. This enables users to focus the view on a particular set of classifications.

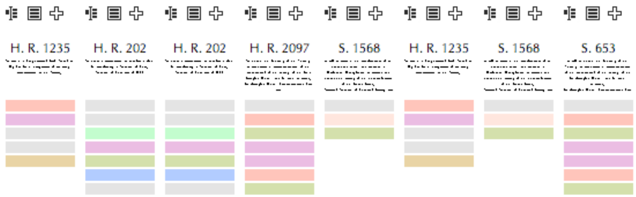

Figure 4 - Minified Mode. Blocks are set to a fixed height thus making long bills easier to scan.

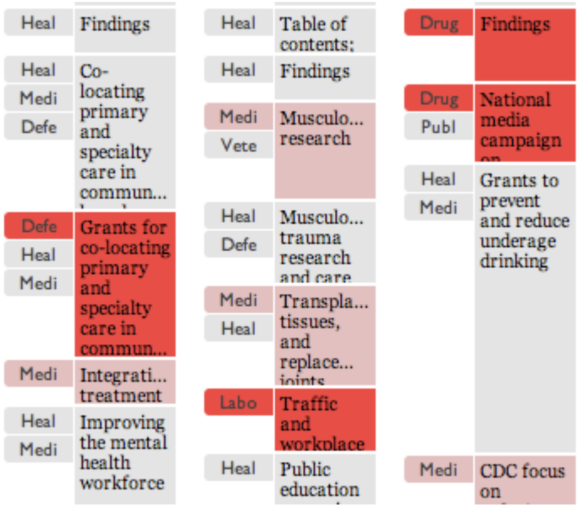
Heal
Findings
Heal
Medi
Defe
Co-locating primary and specialty care in commun...
Defe
Heal
Medi
Grants for co-locating primary and specialty care in commun...
Medi
Integrati... treatment
Heal
Medi
Improving the mental health workforce
Heal
Table of contents:
Heal
Findings
Medi
Vete
Musculo... research
Heal
Defe
Musculo... trauma research and care
Medi
Heal
Transpla... tissues, and replace... joints
Labo
Traffic and workplace
Heal
Public education
Drug
Findings
Drug
Publ
National media campaign on
Heal
Medi
Grants to prevent and reduce underage drinking
Medi
CDC focus on

Figure 5 - Improbability Mode. Red blocks represent sections that are less likely to appear in the bill. Grey colored blocks are unsurprising in the context of the bill.

The visualization also provides a number of different view options. In minified mode (figure 4), each section has a fixed height, enabling users to compare patterns of subjects across bills of greatly varying length. Another view option highlights sections whose probability of appearing in the given bill is low. To calculate these probabilities we measure how often a particular subject (as assigned to individual sections) appears in bills with the same top subject. Using a simple grey-to-red gradient scale, we highlight sections with a lower probability by using a brighter color (figure 5).

Website

The DocBlocks visualization is embedded in a website that enables users to search for and visualize arbitrary collections. Our intent is to enable users to create their own collections that can then be embedded in other social media and shared using static urls.

The website will be seeded with collections of bills that are grouped by their top subject. Users may also search the site using a faceted search interface that includes subject, author, date, and full text. The search results appear as a temporary collection displayed in the same visualization as any other collection.

Implementation

The DocBlocks visualization is implemented using a combination of Javascript and HTML, while the website back-end is built upon Ruby on Rails and PostgreSQL.

Future Work

In our evaluation of DocBlocks, we seek to validate three aspects of the visualization and website. First, the accuracy of our data analysis. Second, the utility of the visualization for exploration of our dataset. And third, the utility of the visualization and website for exploration, communication, and advocacy on the public web.

We plan to approach the first two questions using in-lab studies. Our target users include individuals and groups at varying levels of experience in reading congressional legislation, including political watchdog and advocacy groups as well as individual bloggers and average citizens who are proficient with the web.

However, we believe that the third question can only be addressed by deploying DocBlocks as a publicly available website and promoting its use. Our goal is to launch the site in Q1 2010, collecting data via logging as well as interviewing participants from its user base. Doing so enables us to evaluate the site in a situated context, building upon our experience designing systems that function as components of broader online communities. Through evaluation and revision, we hope that DocBlocks will prove to be a valuable tool for exploration, advocacy, and discussion of key issues that impact both the US and the global community.

Citations

[1] GovTrack.us. http://www.govtrack.us/

[2] HR627 on thomas.gov. http://thomas.loc.gov/cgi-bin/query/z?c111:hr627.enr:.

[3] MAPLight.org. http://maplight.org/.

[4] SunlightFoundation.com. http://www.sunlightfoundation.com/.

[5] Boguraev, B., Kennedy, C., Bellamy, R., Brawer, S., Wong, Y., and Swartz, J. Dynamic presentation of document content for rapid on-line skimming. Proceedings of the AAAI Symposium on Intelligent Text Summarization, (1998), 118-128.

[6] Danis, C.M., Viégas, F.B., Wattenberg, M., and Kriss, J. Your Place or Mine?: Visualization as a Community Component. Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems, ACM (2008), 275-284.

[7] Eick, S., Steffen, J., and Jr, E.S. Seesoft - A Tool for Visualizing Line Oriented Software Statistics. IEEE Transactions on Software Engineering 18, 11 (1992), 957-968.

[8] Hearst, M.A. TileBars: visualization of term distribution information in full text information access. Proceedings of the SIGCHI conference on Human factors in computing systems, ACM Press/ Addison-Wesley Publishing Co. (1995), 59-66.

[9] McCallum, A.K. MALLET: A Machine Learning for Language Toolkit. 2002. http://mallet.cs.umass.edu.

[10] McKeon, M. Harnessing the Web Information Ecosystem with Wiki-based Visualization Dashboards. IEEE Transactions on Visualization and Computer Graphics 15, 6 (2009), 1081-1088.

[11] Viégas, F., Wattenberg, M., van Ham, F., Kriss, J., and McKeon, M. ManyEyes: a Site for Visualization at Internet Scale. Visualization and Computer Graphics, IEEE Transactions on 13, 6 (2007), 1121-1128.

[12] Wise, J.A., Thomas, J.J., Pennock, K., et al. Visualizing the non-visual: spatial analysis and interaction with information from text documents. Proceedings of the 1995 IEEE Symposium on Information Visualization, IEEE Computer Society (1995), 51.