Many Bills: Visualizing the Anatomy of Congressional Legislation

Elif Aktolga
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts Amherst
Amherst, Massachusetts
elif@cs.umass.edu

Irene Ros, Yannick Assogba, Joan DiMicco
IBM Watson Research Center
Visual Communication Lab
Center for Social Software
Cambridge, Massachusetts
{iros, yannick, joan.dimicco}@us.ibm.com

Abstract

US Federal Legislation is a common subject of discussion and advocacy on the web. The contents of bills present a significant challenge to both experts and average citizens due to their length and complex legal language. To make bills more accessible to the general public, we present Many Bills: a web-based visualization prototype that reveals the underlying semantics of a bill. We classify the sections of a bill into topics and visualize them using different colors. Further, using information retrieval techniques, we locate sections that don't seem to fit with the overall topic of the bill. To highlight outliers in our 'misfit mode', we visualize them in red, which builds a contrast against the remaining gray sections. Both topic and misfit visualizations provide an overview and detail view of bills, enabling users to read individual sections of a bill and compare topic patterns across multiple bills. We obtained initial user feedback and continue collecting label corrections from users through the interface.

Introduction

Reading legislative documents is often tedious and can be demanding to both the expert and average citizen. Proposed legislation in the US, in the form of bills, is written in complex legal language, which makes reading a challenging task. Organizations such as OpenCongress (2009) seek to help the average citizen navigate the complexity of US Federal legislation by making the text available online.

The problem of understanding the structure of a bill is particularly acute in long bills, in which case it would be very helpful for the reader to know exactly where to look for parts of interest within a bill. One way IBM Many Bills supports this is by labeling individual sections with their main topic and color coding these topics to make them easy to identify within a bill. Another region of interest for users such as journalists, watch dog groups and concerned citizens are sections whose content differs from the main topic of the bill. This may indicate that the section is worth paying more attention to.

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 – so-called outliers or 'misfits'. One of our favorite examples appears in H.R. 627, shown in Figure 1: The Credit Card Accountability Responsibility and Disclosure (CARD) Act of 2009, which imposes transparency and disclosure requirements on credit card companies. In this figure, each colored rectangular bar represents a section, the smallest unit of a bill. Different colors indicate different topics. This bill is categorized as being about 'Finance' by the Congressional Research Service. However the bill contains Section 512 titled 'Protecting Americans From Violent Crime', which guarantees the right of citizens to carry firearms in National Parks and Wildlife Refuges. As this example shows, even good document-level meta data can hide important aspects of the underlying text, which is a common problem in US Congressional bills. We believe that a modest first step towards addressing this issue is to provide the public with better tools for understanding the complex subject matter of legislation, a core part of the political process. In doing so, we hope to progress towards the goal of 'Total Political Transparency' that Munzner proposed as a Grand Challenge in Information Visualization (2008) – making the massive amounts of data that connect money, politics, and advocacy accessible and understandable to the general public.

Our contribution to this effort is Many Bills: a visualization and website that enables anyone to search and explore the content of the bills brought before the United States Congress. Each bill is presented in reduced form as a column of blocks that preserve relative length; each section of the bill is assigned a color corresponding to its general topic based on the results of our analysis (Figure 1). Users may click on sections to expand and read them, and view bills in minified mode (where each section is equally sized) or in misfit mode, in which outlier sections are colored red, contrasting them against the remaining gray neutral sections.

This paper is organized as follows: in the next section we discuss previous related work, which is followed by implementation details of the website. This includes a description of how we accomplish section topic classification and outlier section detection. We then describe our visualization design, which enables users to view information about the topics or outliers of a bill at various levels of detail. We conclude by presenting initial user feedback and observations that illustrate the strengths of this visualization method, followed by
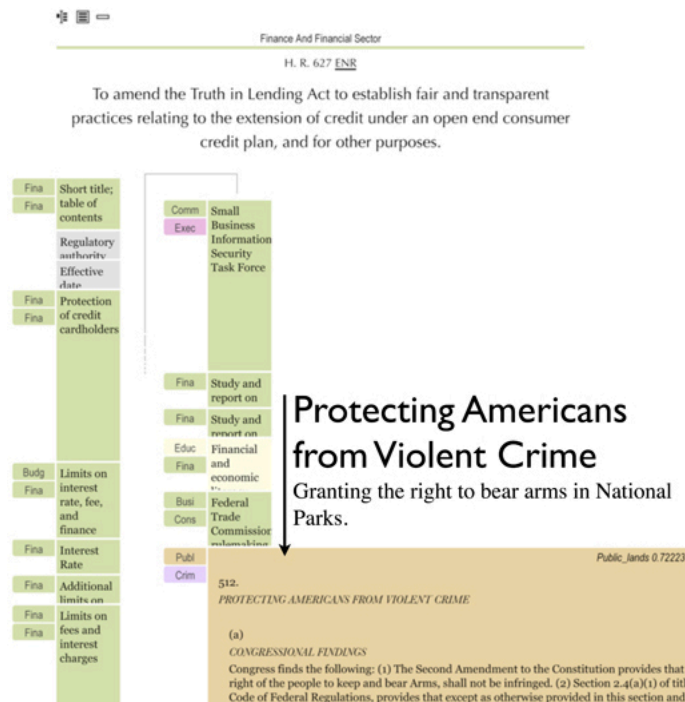
Figure 1: H.R. 627: Example of a bill about credit cards with an outlier section about gun control.

Related Work

There has been much prior work providing information about individual documents. Some, such as Hearst's Tile-Bars (1995), present condensed representations of documents that preserve relative length, indicating points of interest (such as search term hits) within the document using color or other highlighting techniques. Other methods generate semantic, fixed-size thumbnails of documents (Berkner, Schwartz, and Marle 2004; Strobelt et al. 2009). While these methods do provide high-level summary information on individual documents, for our purposes they do not adequately illustrate document structure.

Some of these visualization methods integrate full-text browsing – (Eick, Steffen, and Sumner 1992) display source code files as an aligned set of columns, coloring rows of pixels in each column based on statistics related to the corresponding line of code (such as author or date added). Users may read the lines of code in a popover window by dragging a magnifying box over the display. Similarly, (Boguraev and Bellamy 1998)'s ViewTool combines a high level, topic-based single document overview and a full-text view into a single 3-pane interface. These approaches are close to the model that we wish to support; however we believe it is valuable to show full text detail in the context of the document overview. In addition, Seesoft's focus on line level meta data differs from our goal of providing section level topical overviews derived from the text itself.

Interfaces that use visualization to assist human navigation of the results of automated analysis of text have proved to be valuable tools for gaining a high-level understanding of documents. (Plaisant et al. 2006)'s visualization of Emily Dickenson's correspondence uses a human-assisted machine classification system to assign documents into one of two categories for further analysis. However, its heavy reliance on experts for both classification and interpretation, as well as the low granularity of its categories, limit its application to the more general problem we are attempting to address. (Liu, Selker, and Lieberman 2003) use a classifier to determine the emotional content of text passages and present a visual summary and navigation interface for a document based on those classifications. This work is focused on a navigation of a single document; it does not show a summary of thematic content of a document in context with a collection of other documents.

Implementation

Our goal of supporting government transparency across a broad audience led us to chose the web as our platform for delivery. ManyBills is available as a website at http://manybills.us/ and is built on standard web technologies such as HTML, CSS and Javascript for the front-end and Ruby on Rails for the back-end.

Determining the Topics of Sections

Congressional bills are structured hierarchically consisting of discrete units known as sections that typically cover a single provision of the legislation. Each bill comes with meta data such as the top category assigned to a bill (e.g. 'taxation'), further categories (e.g. 'small business', 'income tax credits'), and the section texts. Whereas the categories at the

bill level are manually assigned by humans from the Congressional Research Service (CRS), there are no classifications available at the section level. Therefore, we determine the topics of sections automatically with a trained classifier: we train a multi-class maximum entropy document classifier for 83 of the classes that the CRS assigns as top categories to bills. The classifier is trained over 59552 bills from the past 9 years of congress with 10-fold cross validation. The Mallet (2002) toolkit is used to train the classifier and then applied to the sections of bills from the 111th congress (2009-2010) to generate a topic for each section of a bill, averaging 90% accuracy during evaluation. More details on this classification can be found in (Aktolga, Ros, and Assogba 2011; Assogba et al. 2011).

Detecting Outlier Sections

Beyond detecting the topical structure of bills, we wanted to further provide readers with insight into bill content by highlighting when a bill section significantly deviates to a topic unrelated to the other topics within a bill. For example, a finance bill with a section on student loans may be topically categorized as education, but is still quite related to its parent topic of finance, whereas a section on gun control in this finance bill is a strong topic deviation, or in our terminology an 'outlier'. In this section we describe how outlier sections can be detected within bills, and used for display in the Many Bills' misfit mode. The approaches described here are detailed in full length in (Aktolga, Ros, and Assogba 2011).

Unlike the example shown in Figure 1, most bills are consistent in their content and may contain minor topic shifts, or what we identify as mild outliers. We aim at identifying two types of outliers – mild and strong. Whether a section is a mild or strong outlier depends on how important (and how dramatic) its topic shifts in relation to the rest of that bill, as determined by the reader. To illustrate this using the H.R. 627 credit card bill described above, a section within it titled 'Privacy Protection for College Students' could be considered a mild outlier or a topic shift, because it still discusses matters related to credit cards and finance, but focuses on a specific target group – college students. This section would not be nearly as interesting to a user looking for anomalies as the one about gun control (Section 512), but still may be worth noting to some readers. We conclude that determining whether a section is an outlier is a challenging task because the decision is somewhat subjective, and depends on the context the section appears in. Another issue is that we want to control the number of falsely identified outlier sections shown to the user. We do not want users to lose trust in the predictions of our algorithms, therefore we aim at high precision and low recall.

A major challenge in this task is that there is no pre-existing truth data for outliers in US Congressional Legislation. Since outliers are contextual to the bills they appear in, we take a generative approach to detecting outliers within bills. We create language models from different units of a bill, such as its individual sections, the entire bill text, or even from all the bills within a category, such as all 'Finance' bills. Then we apply dissimilarity measures to compare the section language models of a bill to other language models built from the different units just discussed, from which we obtain a ranking of the sections indicating the probability of each section being an outlier. We apply a threshold to this outlier score for the final decision of whether the section is marked as an outlier or not (currently 70%).

The most promising techniques use a 2-step approach: the main topical sections within a bill are first detected by ranking the top most relevant sections with respect to the bill's title keywords using Okapi BM25. For this, language models Θ_{S_i} are inferred for each section $s_i \in D$, where D is a bill. We take the top m sections in D from the Okapi BM25 ranking and call them 'main sections'. These are compared to the remaining $k = n-m$ sections for determining the outlier probability of each of the k sections. For this we need to know how each of the k sections compare to the m sections *on average*:

$$P_{\text{outlier}}(s_{\text{LM}}|\Theta_{S*}) = \frac{\sum_{i=1}^{m} P_{\text{dsm}}(s_{\text{LM}}|\Theta_{S_i})}{m} \qquad (1)$$

where s_{LM} is one of the k remaining sections and Θ_{S*} denotes the average section model built from the main m sections only. $P_{\text{dsm}}(s_{\text{LM}}|\Theta_{S_i})$ denotes a single comparison between a section $s \in D$ and s_i:

$$P_{\text{dsm}}(s_{\text{LM}}|\Theta_{S_i}) = \text{dissim}_{\text{norm}}(s_{\text{LM}}||\Theta_{S_i}) \qquad (2)$$

where s_{LM} is the language model inferred from s. This equation denotes the probability of how dissimilar s_{LM} is from Θ_{S_i}. As dissimilarity measures we tried various ones based on Kullback-Leibler Divergence (1951; 1991). One of them that particularly helps with strong outliers is Kullback-Leibler Divergence Contribution (KLC), which we introduce based on inspiration from Lawrie and Croft's work (2003):

$$KLC(P||Q) = \sum_{i \in P \wedge Q} |P(i) \cdot \log_2 \frac{P(i)}{Q(i)}| \qquad (3)$$

where P and Q are two language models built from sections with Q being the base model, and i are terms from sections whose frequencies are compared to each other. These dissimilarity measures aid in distinguishing between outlier and non-outlier sections because they capture the following: (1) terms occurring in potential outlier sections that do not occur in main sections contribute to the outlier score with a positive KL Divergence value; (2) important topical terms contained in main sections that are less frequent in outlier sections contribute to the outlier score with a negative KL Divergence value.

Table 1 shows the ranking of terms in the H.R. 627 bill after applying the KLC measure between the main and remaining sections in this bill: we can see that terms such as 'credit', 'card', and 'consumer' with high contributions towards the KLC score are topical terms, whereas those with low contributions such as 'arms', 'bear', 'fish' rather diverge from the main topic of the bill. By utilizing this information in different ways, some dissimilarity metrics are more successful in catching strong outlier sections efficiently, whereas others do better on milder outliers.
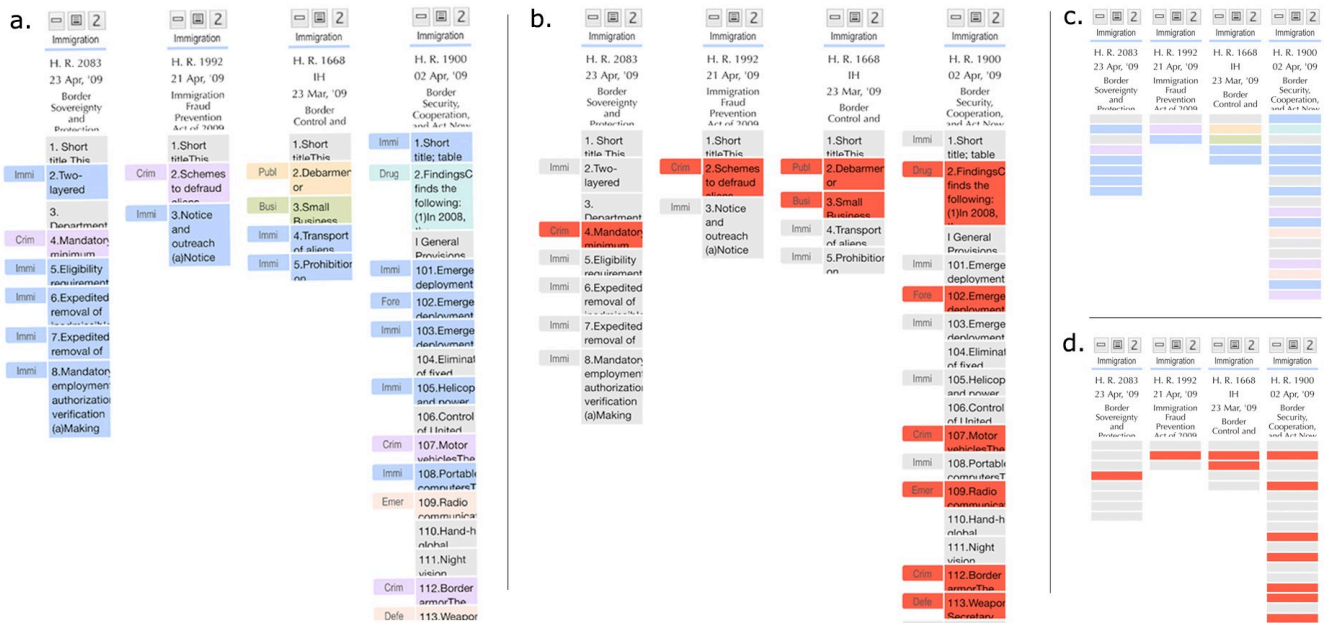
Figure 2: H.R. 627: Bills with their colored sections in normal (a), misfit (b), and minified mode (c with normal, and d with misfits).

tokens	sum of contributions
arms	0.099
bear	0.099
firearms	0.099
fish	0.099
land	0.099
protecting	0.099
. . .	. . .
payment	8.3
account	8.7
fees	9.3
consumer	19.9
card	24.8
credit	30.3

Table 1: Analysis of token contributions towards the KLC score in all the sections of the credit card bill from Figure 1.

Visualization Design

Our visualization is designed around various aims, two of which result from the features just described in the previous sections: visualizing the topics of sections and presenting outlier sections versus non-outlier (neutral) sections in an appropriate manner to users. We have also started collecting user feedback on the website to improve the quality of our topic classifications and outlier detections. Each section can be flagged as having a correct or incorrect topic classification or outlier label. For both visualizations, view modes are intended to be combined; one may view bills in minified or full mode while examining sections, as shown in Figure 2 (a and b versus c and d).

Visualizing Section Topics

Given that we have calculated the topic classifications for the sections in a bill, our aim is to present this topical structure in a manner that can be consumed at a glance. Although we have merged similar topics assigned by CRS to yield a total of 83 reduced topics, this is still a large number, which restricts our choices in distinguishable colors for the section topics. Another visualization issue is the text of the sections. A strong goal of our design is to stay true to the text of the bills while providing abstractions that make them more easily consumable. As evident in Figure 2 (a and c), we present each bill as a series of section bars arranged vertically. The height of a bar maps linearly to the length of the text of the section in the bill. We then assign a color to each block based on its classifications. The subject with the highest confidence score is chosen, and then we use the color of its group as the background color of the section. Sections whose highest-scoring subject's confidence is below 10% are rendered in gray.

By grouping several subjects into a single color, it has become harder to discern a specific subject through color alone. To remedy this, we present small subject badges as side tabs appearing to the left of a section. The badges contain the first 4 letters of the subject they represent. Upon mousing over, users can see the full subject name alongside the confidence score associated with it in a tool tip. To provide users a glimpse of the section's content, we have added a small preview snippet describing the contents of the section. We use the section's title or the first 100 characters of the text to obtain the snippet.

With over 14,000 bills in our corpus, displaying them all at once is impractical. Thus, our interface presents bills in

collections, a grouping of one or more bills. In a collection, each bill is represented as a single column. Collections scroll left to right, if need be. To ensure browser performance, we only display 50 bills per page; admittedly, this has the drawback of requiring users to page through large collections. We seed the site with collections based on groupings of bills by their top subject. Additionally, users can group bills as they see fit into their own collections.

Visualizing Outlier Sections

Outlier or misfit sections can be viewed in the 'Misfit Mode'. An example is shown in Figure 2 (b and d). This mode attempts to reveal sections that do not fit with the general topic of the bill using three colors: a section is colored in red if its probability of being an outlier is above the outlier threshold of 70%; it is colored gray if it is not an outlier, and it is colored in orange if users submitted conflicting feedback about this section. This means that at least two users have different opinions about the section in question. This spare choice of colors allows users to directly distinguish outlier sections from others, which is why we do not include the colorful topic section classifications in this mode. Our aim is to point users to interesting sections, and this visualization allows us to draw their attention towards them immediately.

Evaluation and User Feedback

Several examples have provided validation during our work on Many Bills; we believe these show various ways in which the tool can help users find interesting information quickly within a bill.

Given our visualization, several observations are easily made within minutes: the controversial H.R. 4872 Reconciliation Act of 2010 makes substantial changes to health care regulation, while also containing substantial portions of education legislation. This is easily observed by turning off all colors except for those related to education. Instantly, a portion at the end of the bill is lit up, revealing the exact location of these education sections. During our initial evaluation period, our users made the same discovery simply by scanning the colors present in the bill while looking at it in minified mode. We believe these dense documents can be made more digestible by providing such visual segmentation of a bill into constituent topical clusters.

We also sought the opinion of four users who all had previous experience dealing with legislation in a professional capacity. These subject matter experts used Many Bills for periods ranging from 30 minutes to 2 hours, after which we interviewed them over the phone to get their impressions and thoughts.

Our users all appreciated that Many Bills presents bills without altering the original text while providing a layer of abstraction that includes visual guides as to where to look. While this certainly still leaves the challenge of visualizing the dense language of individual sections, our testers all agreed that our visualization made it easier to narrow into portions of a bill that an individual would find interesting. The most obvious advantage our users perceived was the quick overview of a bill's subjects provided by our visual-ization. Several participants confirmed that the visual segmentation created by our coloring of sections helped them navigate the bills and make decisions about what to read next.

Two of our users expressed a preference for browsing the collections in the minified or misfit modes. This allows users to get a quick overview of the topics in a bill and browse by color, often picking out things that stand out. Our users also confirmed that the combined usage of the misfit mode with the minified mode is useful.

Two users indicated that they were 'overwhelmed' by the interface at first and that there were 'too many things on the screen' at once. However, after some time users became accustomed to the interface and appreciated the depth of exploration available to them. Our choice of using both horizontal scrolling and paging confused most users, who were unsure as to why we had both instead of only one of them. Additionally, our extended search and filtering options were appealing but too hidden. Several users commented that having to go to a collection view in order to perform more complex search queries was unintuitive. Two users also expressed their desire to have a way to compare the bills on a more granular level, much like a diff tool would compare two documents.

Overall, our users felt that Many Bills is a valuable addition to the collection of available tools. We have started gathering feedback from the general population as well to better cater to lay users, who were not a part of this initial feedback iteration.

Conclusions and Future Work

In this paper, we have described the goals, design, and implementation of Many Bills, a visualization and website that provides rich interactive views of the topic structure of US Congressional Legislation to the general public. Our initial user evaluations indicate that this is a valuable tool for helping people understand the content of bills. We realize that it is important to appropriately visualize information retrieval results in order to make them comprehensible to end users. Traditionally, retrieval results are used to measure the efficiency of a system, but here we tried to bridge the gap between the data and end users by means of visualization. Color is particularly valuable for highlighting extremes in data, which we have employed in the misfit mode.

As for future work, a major aim already in progress is collecting user feedback in the form of labels for both the topic classification and outlier detection of sections. This will greatly help us in improving and extending the current techniques.

We might also explore other forms of visualizations for bills. For example, navigating the relationships between bills was a frequently requested feature by our users. Further, a view that shows differences between bill versions, or markup that reveals detected entities (such as persons, places, and dollar amounts), and alternative summary views that feature more detail than the current display affords may be useful extensions. Experience indicates that our users will continue to provide plenty of suggestions as the public instance of this project matures.

References

Aktolga, E.; Ros, I.; and Assogba, Y. 2011. Detecting Outlier Sections in US Congressional Legislation. In *Proc. 34th annual international ACM SIGIR conference on Research and development in informaion retrieval*. New York, NY, USA: ACM.

Assogba, Y.; Ros, I.; DiMicco, J.; and McKeon, M. 2011. Many Bills: Engaging Citizens through Visualizations of Congressional Legislation. In *Proc. 2011 annual conference on Human factors in computing systems*, CHI '11, 433–442. New York, NY, USA: ACM.

Berkner, K.; Schwartz, E.; and Marle, C. 2004. SmartNails - Image and Display Dependent Thumbnails. In *Proceedings of SPIE*, volume 5296, 53–65.

Boguraev, B., and Bellamy, R. 1998. Dynamic Document Presentation. *SIGCHI Bull.* 30:72–77.

Cover, T. M., and Thomas, J. A. 1991. *Elements of Information Theory*. Wiley series in Telecommunications. New York, NY [u.a.]: Wiley.

Eick, S. G.; Steffen, J. L.; and Sumner, Jr., E. E. 1992. Seesoft-A Tool for Visualizing Line Oriented Software Statistics. *IEEE Trans. Softw. Eng.* 18:957–968.

Hearst, M. A. 1995. TileBars: Visualization of Term Distribution Information in Full Text Information Access. In *Proc. SIGCHI conference on Human factors in computing systems*, CHI '95, 59–66. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co.

Keim, D., and Munzner, T. 2008. Grand Challenges for Information Visualization. Presented at IEEE InfoVis, Columbus, Ohio, USA.

Kullback, S., and Leibler, R. A. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics* 22(1):79–86.

Lawrie, D. J., and Croft, W. B. 2003. Generating Hierarchical Summaries for Web Searches. In *Proc. 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 457–458. New York, NY, USA: ACM.

Liu, H.; Selker, T.; and Lieberman, H. 2003. Visualizing the Affective Structure of a Text Document. In *CHI '03 extended abstracts on Human factors in computing systems*, CHI EA '03, 740–741. New York, NY, USA: ACM.

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

Plaisant, C.; Rose, J.; Yu, B.; Auvil, L.; Kirschenbaum, M. G.; Smith, M. N.; Clement, T.; and Lord, G. 2006. Exploring Erotics in Emily Dickinson's Correspondence with Text Mining and Visual Interfaces. In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '06, 141–150. New York, NY, USA: ACM.

Strobelt, H.; Oelke, D.; Rohrdantz, C.; Stoffel, A.; Keim, D. A.; and Deussen, O. 2009. Document Cards: A Top Trumps Visualization for Documents. *IEEE Transactions on Visualization and Computer Graphics* 15:1145–1152.

2009. US Congressional Legislation. Available at GovTrack:. http://www.govtrack.us/data/us/111/.