Kongress: A Search and Data Mining Application for U.S. Congressional Voting and Twitter Data

Douglas Grosvenor, Jeffrey Kendall, Amy Sanders, Chang-Tien Lu
Computer Science Department, Virginia Tech
7054 Haycock Road, Falls Church, VA, 22043
{drg4m, jdk34, abs5055, ctlu}@vt.edu

ABSTRACT

As the world braces for the impact of the sequestration, international conflicts, and other decisions facing the US congress, people are wondering how their congressmen's decisions will affect their lives. Traditionally, to understand what issues a congressman found import, interested constituents would synthesize voting records, bills, and other disparate data sets to understand their congressman's habits. Fortunately, technology can now be used to integrate and display this information in an informative and visually appealing way. In response to this need to understand the behavior of congressmen, we have developed a mobile-based search and data mining application that provides users with the ability to analyze a large amount of social media data from Twitter, as well as data from the United States Congressional voting records. The application is focused on identifying patterns, anomalies, and associations between members of congress and external users to determine influential users within and outside Congress. This paper introduces the motivation behind the application – Kongress - and then progresses into the system architecture. The applications features include the ability to search congressional tweets, votes, and bills, and a geospatial visualization of congressional tweets. We also demonstrate how a user could use Kongress to understand the motivation behind a congressman' decisions.

Categories and Subject Descriptors

H.3.0 [**General**]: Information Storage and Retrieval; H.5.2 [**User Interfaces**]: Graphical User Interfaces

General Terms

Human Factors, Measurement

Keywords

social media data mining, us congress, influential users, mobile information retrieval, spatio-temporal search engine

1. INTRODUCTION

Social media websites provide an ever-evolving and rich set of data to analyze for patterns, anomalies, and associations. Facebook, for example, surpassed the one billion-user mark in October 2012 and has over 140.3 billion friend connections, which can be analyzed for social patterns [1]. Twitter users currently tweet approximately half a billion times each day, with most of those tweets sent from mobile phones that contain a global position service (GPS) [2].

Because of this, significant attention has been directed towards efficiently analyzing this data for information and judgments relevant to consumers [3, 4, 5]. Some of the hard problems identified in social media data analysis include extracting relevant location information from unstructured text data and structured metadata, identifying themes in unstructured text data (i.e., topic modeling), and determining relevant associations among users that can help a certain consumer make informed decisions [6].

The application described in this paper - Kongress - will not address consumers, but rather constituents. Many congressmen have joined the 'Twitterverse' and are frequently tweeting about their most important issues. Kongress attempts to harness this and enable the user (constituents or any global users interested in congressional issues) to understand who in congress is influential for a user-specified topic. In addition to enabling Kongress users to understand which congressional members are the most influential based on Twitter data, Kongress also enables users to understand which congressional users are the most influential based on congressional voting records.

The main contributions of Kongress are as follows:

- **Mobile Search Platform**: users are able to query vast amounts of data related to congress, including congressional voting records and congressmen's tweets, via the users' mobile devices.
- **Keyword Map Overlay**: users can plot tweets on a map that match a user's search criteria to see which congressmen are tweeting about a specific topic; this gives users a geospatial perspective to the different topics congressmen are concerned with
- **Novel Congressional Analytics**: with the use of open source technologies, users are able to view and analyze congress' voting records. This analysis allows the user to see if there are any influential people within congress, or on Twitter, based upon congressmen's connections on Twitter (friends, followers)

2. SYSTEM ARCHITECURE OVERVIEW

To provide the necessary functionality for its users, Kongress needs instant access to the data, as well as a sleek and intuitive user interface (UI) to display the queried information. Because of this, several open source technologies were used to harness and display the Twitter and voting record data

SIGSPATIAL'13, Nov 05-08 2013, Orlando, FL, USA
ACM 978-1-4503-2521-9/13/11.
http://dx.doi.org/10.1145/2525314.2525328

Starting with the web tier, users access Kongress through an Nginx reverse proxy, which provides a unified access point to the Kongress web application. After accessing the Nginx reverse proxy, users connect to a barebones Jetty webserver that hosts the main application. This Jetty webserver was chosen other webservers because Kongress only needs the basic Java servlet functionality and the Jetty webserver provides better reliability and performance than other webservers. Also, the only portion of Kongress that needs the Java servlet functionality is the analytic features, which will be discussed in Section 3.7.

At this point in loading the application, the user will start seeing Kongress's UI, which was built with Sencha's easy to use and flexible mobile framework, Sencha Touch. Sencha Touch is a web-based cross-platform UI that enables development for mobile devices (e.g. iPhone, iPad, and Andriod), as well as for desktop/laptop computers. It is a JavaScript/HTML5 framework based on EXT-JS [11]. We chose Sencha Touch over other HTML5 frameworks because the touch gestures are smoother which enhance the overall usability of the application. As the user queries Kongress, searches are executed against the searching and indexing portion of Kongress, hosted by ElasticSearch. The searching and indexing features are covered in Section 3.6.

3. FEATURES

The following section outlines the capabilities of the application and different analyses a user can perform. Since Kongress is a mobile application, there are different "pages," a total of five to be exact, where the user can utilize various analytical features.

3.1 The Search Page

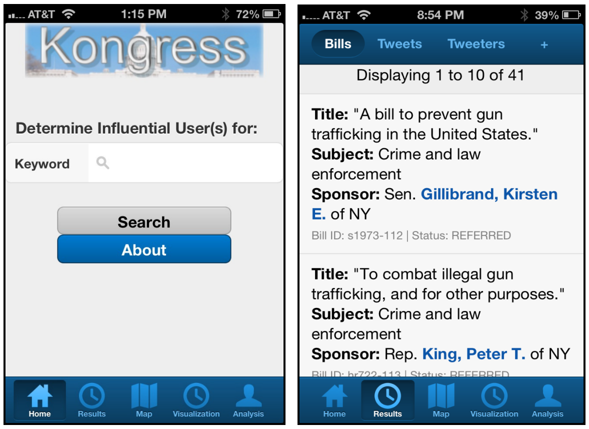
After starting application, the user is greeted by Kongress' Google-like landing page, which can be seen in Figure 1 (left image). There is a simple text box to enter a search query, and two buttons, one to execute a search and the other to find out more information about the application. This streamlined user interface allows the user to quickly understand the concept of the application and promptly use the application by entering a search query to find the most influential congressmen on a particular topic. When entering a search query, that search is executed against all of the tweets, as well as all of the congressional bills that were indexed into Kongress' search engine. If multiple terms are in the user's search query, the search engine uses a Boolean AND by default but the Search Page can take in various Boolean operations and other matching keywords.

3.2 The Textual Results Page

Once the user inputs a search query and taps the search button, the user is taken to the textual results page, which will list the results in one of three ways:

3.2.1 Bills

The list of bills with the related search term is displayed, as seen in Figure 1 (right image), with information that includes the official title, the main subject of the bill, the congressmen that nominated the bill, the bill's id, and the status of the bill. When the user taps on a specific bill, a shadowbox will popup and display more information, including information from the results page plus when the bill was introduced, the last time the bill was updated, and the summary of the bill.



Determine Influential User(s) for:

Displaying 1 to 10 of 41

Title: "A bill to prevent gun trafficking in the United States."
Subject: Crime and law enforcement
Sponsor: Sen. **Gillibrand, Kirsten E.** of NY
Bill ID: s1973-112 | Status: REFERRED

Title: "To combat illegal gun trafficking, and for other purposes."
Subject: Crime and law enforcement
Sponsor: Rep. **King, Peter T.** of NY
Bill ID: hr722-113 | Status: REFERRED

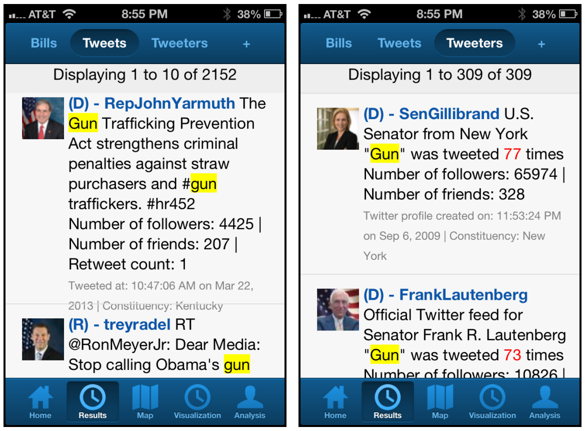
Figure 1: Kongress Home Search Page (left image) shown on an iPhone and Textual Results Page (right image) showing bills containing the search term "gun" on the iPhone

3.2.2 Tweets

The tweets with the related search term are displayed, along with other information such as the congressmen's political party, the congressmen's Twitter handle (e.g., RepJohnYarmuth), the full tweet, with the search term highlighted, the tweeter's number of followers, the tweeter's number of friends, the number of times that tweet was retweeted, the time the user sent that tweet, and where that congressmen represents; this can be seen in Figure 2 (left image).

3.2.3 Tweeters

The list of congressmen who tweeted the specified search query will be displayed in descending order by the number of times they tweeted that search query. Here, Kongress displays the congressman's political party, their Twitter handle (e.g., SenGillibrand), the profile description, the number of times that congressman tweeted the search term, the congressman's number of followers and friends, when the congressman joined Twitter, and the congressman's political district. All of this information can be seen in Figure 2 (right image).



Displaying 1 to 10 of 2152

(D) - RepJohnYarmuth The Gun Trafficking Prevention Act strengthens criminal penalties against straw purchasers and #gun traffickers. #hr452
Number of followers: 4425 | Number of friends: 207 | Retweet count: 1
Tweeted at: 10:47:06 AM on Mar 22, 2013 | Constituency: Kentucky

(R) - treyradel RT @RonMeyerJr: Dear Media: Stop calling Obama's gun

Displaying 1 to 309 of 309

(D) - SenGillibrand U.S. Senator from New York
"Gun" was tweeted 77 times
Number of followers: 65974 | Number of friends: 328
Twitter profile created on: 11:53:24 PM on Sep 6, 2009 | Constituency: New York

(D) - FrankLautenberg Official Twitter feed for Senator Frank R. Lautenberg
"Gun" was tweeted 73 times
Number of followers: 10826 |

Figure 2: Text Results Page (left image) showing Tweets containing the search term "gun" on the iPhone and Textual Results Page (right image) for Tweeters who tweeted the search term "gun" shown on the iPhone

If the user is interested in a particular congressman, the user can click on a tweeter and a shadowbox will appear. This shadowbox displays information about the tweeter, as well as the top five congressmen that have a similar voting record to the selected congressman. This selection is determined by a simple matching coefficient, which is the ratio of the total number of matches to the total number of votes. This matching is done over the entire voting dataset and the total number of congressional tweeters. Along with the top 5 congressmen, this shadowbox also displays the congressman's tweets on that specified query.

Note that these three views give the user a general idea related to which congressmen tweet about a particular topic the most and are likely to be influential on that particular search term or topic; however, this does not necessarily link the Twitter and congressional data to enable the user to identify the most influential users.

3.3 The Map Page

The Map page displays a geographic interface to view the distribution of tweets across the United States as shown in Figure 3. Since none of the tweets had any geographical information, Kongress displays the tweets as if they were originating from the congressman's home district. (In the case of senators, which represent an entire state, a geo-location in the approximate center of the state was used.) This map page contains an overlay of the 2012 congressional districts to show where each congressional boundary is; this overlay can be enabled/disabled by clicking the 'Clear Districts' button. Along with the districts, Kongress displays a basic pin-based reference point to show where the congressional offices are, as well as two different heat maps.

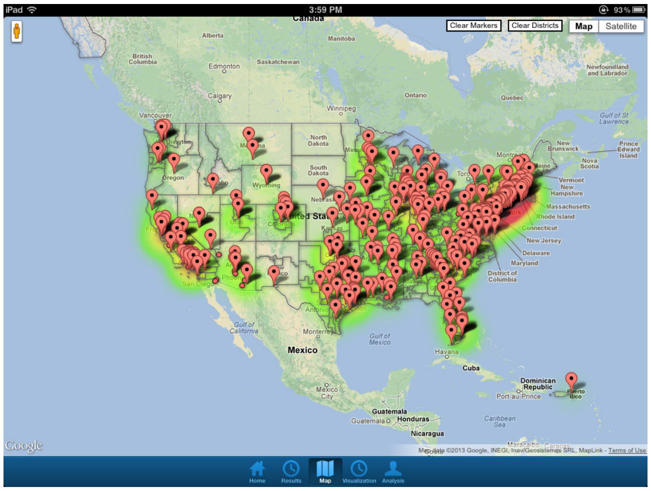


Figure 3: Map Page showing results for the search "gun"

Like the district overlay, the pins and heat maps can be enabled/disabled. The default heat map displays from where the search phrase was tweeted (i.e., the congressmen's home district/state). The other heat map displays the density of the tweets for that particular topic. For example, if a congressman tweeted about guns 25 times, Kongress will weight that location 25 times so it has a higher density than another district/state whose congressman only tweeted about guns once. This allows a user to get a broad understanding of who talks about a specific topic versus which areas/congressmen talk about the same topic the most. All of these overlays, pins, and heat maps were plotted using Google Maps API [9]

3.4 The Visualization Page

The goal of the visualization page is to give the user an alternative way of interacting with the data. To do this, Kongress uses an open-source tool RapidMiner to create visualizations based on voting records and different algorithms to analyze the data [12]. One of the visualizations created can be seen in Figure 4. To get a more in-depth understanding of the specific algorithms used, refer to Section 3.7 Data Analyzer.

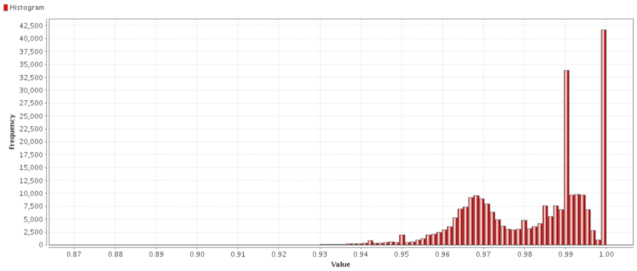


Figure 4: A histogram showing cosine similarities for all congressmen-pair permutations of their voting histories in the 112th and 113th Congressional sessions

3.5 The Influence Page

Along with the data visualizations, Kongress provides topic-specific analysis on which Twitter users (not necessarily congressman) have a potential impact on congressman's decisions. This page takes the search parameter to retrieve all bills and their associated votes, and then analyzes how congressmen's friends could have impacted the congressman's vote. For each 'yes' vote a congressman gave, all of their Twitter friends get a +1 in the rankings. After all of the congressmen are iterated over, the results are ranked in descending order and returned to the user; this is repeated for each vote type (no, not voting, present). Kongress did not index all of the profiles of the congressman's Twitter friends (this would have been around 450,000 more profiles to index), so only the Twitter id is displayed, and the number of "votes" each user got.

3.6 Data Indexer and Searcher

Like most data retrieval applications, Kongress needs a quick and efficient way to search copious amounts of data. There are many open source tools suited for this; one such tool that we used is a relatively new application called ElasticSearch. ElasticSearch is very similar to Apache Solr, but it allows developers to easily ingest new data types without having to create a schema for the data. ElasticSearch creates a schema for each "document type" as it is ingested. Sharding of data in ElasticSearch is also very easy, which adds to its ease of use and performance. Like Solr, ElasticSearch allows users to search the data via RESTful calls, and provides Kongress' user interface, Sencha, an easy access point to the data.

Because congressmen need to uphold a professional appearance, the issue of poorly-spelled/abbreviated tweets is not an issue as it would be if this project mined a different set of social media data. Additionally, every congressman can be traced back to a specific geo-location, specifically their home district or state of representation. This fact allows Kongress to geo-locate every congressional tweet in its dataset, which addresses the hard problem of geo-locating social media. This geo-location conclusion makes the assumption that the content of the tweet is related to the congressmen's home district or state of representation. This assumption is logical because the congressman represents his constituents, which are located in that district and state. Those constituents are the intended users of the Kongress application, who have a goal to determine their congressmen's views on given topics.

Kongress utilizes four main indices to store the separate datasets, including bill index, voting record index, Twitter user index, and tweet index. A fifth index is created so that all of the connections between the congressmen and their followers/friends (network index) could be stored. This was done so that a user could easily and efficiently access a tweeter's data without having to retrieve all of the user's followers/friends (some congressmen had over 500,000 followers). With the RESTful calls, one can put in different parameterization to filter, sort, and query different aspects of the data. Also, considering Twitter is an ever-evolving ecosystem, we decided to pull a static set of information; however, if the need arose, we could dynamically update the five indices. The cutoff date for the data that was loaded into Kongress is April 2013 [13, 14, 15, 16].

3.7 Data Analyzer

Kongress utilizes two automated data analysis features: (1) recommending congressional Twitter accounts to view based off similar voting records and (2) identifying Twitter accounts that are followed by congressmen which could exhibit some degree of influence on congressional voting records

When a user views a congressional member's Twitter profile in Kongress, a variety of information is displayed including five other congressional member's Twitter feeds that exhibit similar voting tendencies to the currently viewed Twitter profile. These recommendations were generated through an analysis of voting records from the 112th and 113th congressional sessions. Each congressman's voting record was compiled and treated as a vector of k length, where k is the number of votes in both sessions of congress. Then all of these vectors were compared iteratively using the Simple Match Coefficient (SMC).

After all congressional pairs are evaluated, the top five congressmen are displayed in the top five most similar congressional Twitter feed. One caveat to this analysis is that only congressmen with Twitter profiles are considered. To measure the SMC, Kongress uses the open-source data-mining tool RapidMiner and the results of this analysis can be seen in Figure 4.

The second analysis identifies Twitter users who have, or could likely have, more influence on congressmen's voting habits for a specific topic (e.g., gun control). This assessment is based on correlations between a Twitter user's friends and followers on Twitter and congressmen's individual voting records. This portion of Kongress queries all bills based upon the users search terms and returns the matching bills and their associated votes, and then analyzes how congressmen's friends could have impacted the congressman's vote. This analysis is done as follows:

Every congressman's friend receives a ranking based upon the total returned votes. This ranking is calculated by assigning points based on the number of "yes" votes a congressman completed in the 112th and 113th Congressional sessions. For each 'yes' vote, all of that congressman's Twitter friends get a single point in their rankings. This is completed for every congressman and every returned bill. After all of the congressmen are iterated over, the resulting Twitter friend scores are ranked in descending order; this is repeated for each vote type (no, not voting, present).

This allows the user to determine which members of Twitter are likely to exert some level of influence on the entire congress's voting habits in the 112th and 113th sessions. An example of this can be seen in Table 1 for the "Yes" and "No" votes for bills with the word "gun" in it.

Table 1. Top ranked influential scores for the query "gun"

Yes	No
politico	davidhawkings
davidhaswkings	SenJohnMcCain
WhiteHouse	politico
washintonpost	SpeakerBoehner
mikeallen	mikeallen
cspan	washingtonpost

4. CONCLUSION

Through the data compiled in Kongress, along with the information retrieval and data mining tools that it offers, we accomplished the goals set out in the beginning of this paper. Kongress successfully bridges two related datasets, congressional Twitter data and congressional voting records, to provide an intuitive application that helps identify trends between congressional Twitter and voting records behavior. Constituents are now able to follow their congressmen with better scrutiny and identify what topics their congressmen hold in high esteem, empowering constituents to elect officials that truly represent them.

5. REFERENCES

[1] http://www.huffingtonpost.com/2012/10/04/facebook-1-billion-users_n_1938675.html

[2] http://news.cnet.com/8301-1023_3-57448388-93/twitter-hits-400-million-tweets-per-day-mostly-mobile/

[3] http://www.forbes.com/sites/onmarketing/2012/06/28/social-media-and-the-big-data-explosion/

[4] http://www.destinationcrm.com/Articles/Editorial/Magazine-Features/Transforming-Social-Media-Data-into-Predictive-Analytics-85687.aspx

[5] http://www.customerthink.com/article/from_big_data_to_big_ decisions _3_ways_analytics_can_improve_retail_experience

[6] http://www.public.asu.edu/~pgundech/book_chapter/smm.pdf

[7] M. Ye, D. Shou, W.-C. Lee, P. Yin, and K. Janowicz, "On the Semantic Annotation of Places in Location-Based Social Networks, " Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ACM KDD 2011), pp. 520-528, San Diego, CA, Aug. 21-24, 2011.

[8] Rui Li, Shengjie Wang, Hongbo Deng, Rui Wang, Kevin Chen-Chuan Chang: Towards social user profiling: unified and discriminative influence model for inferring home locations. KDD 2012:1023-1031

[9] Google Maps API: https://developers.google.com/maps/

[10] http://politics.nytimes.com/congress/

[11] Sencha Touch: http://www.sencha.com/products/touch/

[12] RapidMiner: http://rapid-i.com

[13] GovTrack.us Data (bills and votes): http://www.govtrack.us/developers/data

[14] List of Congressmen who use Twitter: http://twitter.pbworks.com/w/page/1779986/USGovernment

[15] List of Congressmen who use Twitter: http://www.tweetcongress.org/officials/tweeting

[16] Twitter API for Twitter data (user, tweets, friends/followers): https://dev.twitter.com