Content of Municipal Government Websites

Markus Neumann Bruce Desmarais Hanna Wallach

March 22, 2017

Abstract

We study the content of municipal government websites....

1 Introduction

2 Background

Grimmelikhuijsen (2010) run an experiment in which citizens are exposed to randomly selected levels of information about local government council minutes. They find a negative relationship between the information level and perceptions of competence in the local government. This raises an interesting question regarding whether citizens or more likely to participate when they perceive competence or when they perceive incompetence.

Wang, Bretschneider, and Gant (2005) present a widely cited 'model' for evaluating the accessibility of information on government websites. This is an important paper with which we should be familiar at a very detailed level as we use archived web content to assess the volume/accessibility of information provided by local governments.

Osman, Anouze, Irani, Al-Ayoubi, Lee, Balcı, Medeni, and Weerakkody (2014) is less relevant, but they develop a multi-item measure to predict the level of citizen satisfaction with e-government services.

Grimmelikhuijsen and Welch (2012) conduct an enormously relevant study. Insofar as we analyze what predicts openness of government websites, we will be replicating and building upon this study. They focus on Dutch municipal websites, and their approach is fairly limited in scope and highly manual (which we can compliment). For example, one of the dependent variables "Decision-making transparency," is measured "using a discrete (1/0) indicator for whether the underlying principles or reasons for local air pollution policies were given on the Web site."

3 Data

The General Services Administration (GSA) maintains all .gov addresses, and provides a complete[1] list of all such domains to the public through GitHub[2]. This list is updated once per month -

[1] Domains used for testing and internal programs are excluded.

[2] https://github.com/GSA/data/tree/gh-pages/dotgov-domains

we rely on the version released on January 16, 2017. The data from the GSA contains the following variables: One, domain name, specifically, the all-uppercase version of domain and top-level domain (for example, 'ABERDEENMD.GOV'). Two, the type of government entity to which the domain is registered, such as city, county, federal agency, etc. Three, for federal agencies, the name is specfied. Finally, the city in which the domain is registered, is noted.

Here, we focus only on cities. As a first step, we use a webdriver-controlled browser (Firefox/Selenium/Geckodriver) to test whether all of the city websites actually work. Of the 2425 domains listed by the GSA as cities, 292 are not accessible. Furthermore, the .gov domain, as registered at the GSA, is frequently not the website a city actually uses. In many cases, these sites redirect to another address, sometimes not a .gov domain (in this case, we simply use this domain). We record these URLs, as they are required to retrieve the images websites stored in the Wayback Machine (WbM).

In order to provide an overview of our coverage (as not all cities, towns and villages use .gov addresses), we merge this list with U.S. Census data[3]. Here, several limitations in the GSA data need to be accounted for: One, even though the GSA nominally separates websites of cities and counties, some of the domains categorized as cities actually belong to counties. The same is true for townships and boroughs. Ergo, we eliminate all websites belonging to these three types of entities by hand. Furthermore, the city name, as given by the GSA, refers to the city in which the domain is registered, which is not necessarily equivalent to the city the website serves. In many cases, a website of a larger city may be registered to one of its subdivisions (for example, the website of New York is registered to Brooklyn), or vice versa (for example, the website of Homecroftin, a small town within Indianapolis, is registered to the city as a whole). Consequently we fix mismatches between websites and cities manually. Finally, a number of cities are simply misspelled, which we also correct by hand.

After the counties, townships and cities that cannot be matched to the Census data[4] and duplicate websites (some cities have more than one website) are removed, 1813 domains/cities remain.

These cities contain 90,616,865 people, and thus about 28% of the U.S. population (see figure 1).

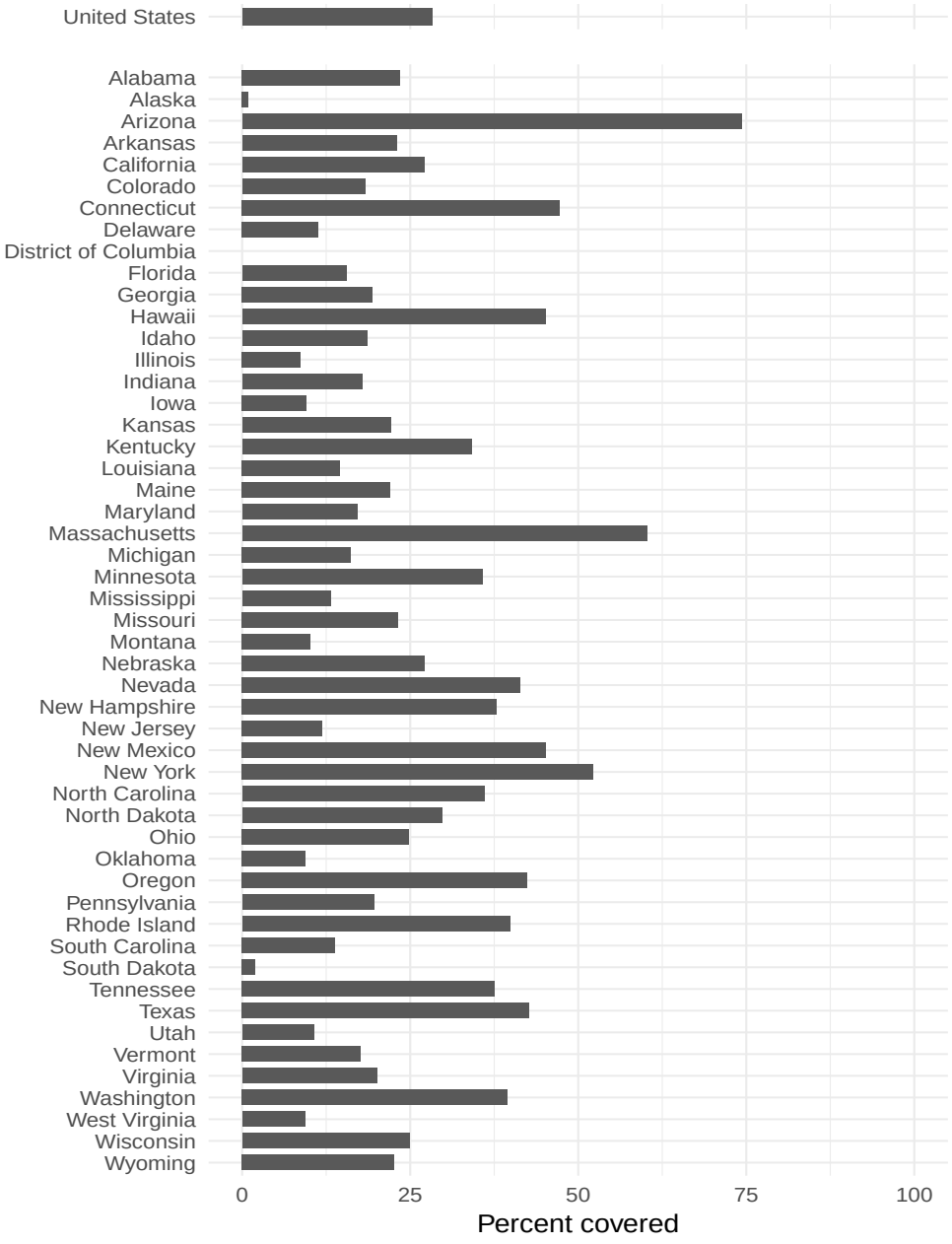
We use the resulting list of websites to acccess their copies stored in the Internet Archive's Wayback Machine. To this end, we rely on the Ruby Gem 'Wayback Machine Downloader'[5] (WbMD). We supply the URL that each .gov website redirects to to the WbMD, which then downloads every file present in the WbM from a snapshot in October 2016, or, if not available, as soon as possible after this point.

<Note: We have not actually done this last step for all websites (however, the R script which runs the Ruby package is already set up to do so once we need to). Instead 10 websites were randomly sampled from an older version of the GSA list, which still contained counties and townships, which is why one of the 10 websites is from Dutchess County, NY.>

[3] http://www2.census.gov/programs-surveys/popest/datasets/2010-2015/cities/totals/sub-est2015_all.csv

[4] There are five cities that are not contained in the Census data

[5] https://github.com/hartator/wayback-machine-downloader


United States
Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
District of Columbia
Florida
Georgia
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
Missouri
Montana
Nebraska
Nevada
New Hampshire
New Jersey
New Mexico
New York
North Carolina
North Dakota
Ohio
Oklahoma
Oregon
Pennsylvania
Rhode Island
South Carolina
South Dakota
Tennessee
Texas
Utah
Vermont
Virginia
Washington
West Virginia
Wisconsin
Wyoming
0
25
50
75
100
Percent covered

3.1 California City Websites

It would be fine to focus on California as a case. First, we need to answer some preliminary questions about the data.

1. For what percentage and number of CA cities can we find data from the WBM?
2. For how many election cycles can we find political leadership data for these matched cities?
3. CA is a deep blue state, in what number and percentage of cities is the local leadership majority Republican?
4. Relatedly, in a typical election cycle, for how many cities do we see a transition in party leadership (i.e., a shift from majority D (R) to majority (R) D).

3.2 Topic modeling

We hypothesize that a change in leadership from one party to the other will lead to a change in website content because the two parties have different agendas. Democrats have a predilection towards policies that promote social and economic equality, whereas Republicans like to emphasize small government as well as law and order. Documents uploaded to city websites are expected to be a reflection of these preferences. The author-topic model (Rosen-Zvi, Griffiths, Steyvers, and Smyth 2004) captures this concept quite well, because it assumes that different authors prefer different topics, each of which is associated with a group of words. We classify all documents that have been present before the change in party control as from one party[6], and the new documents as stemming from party 2. The number of topics *j* would be set to at least three, where one topic would (hopefully) be reserved for purely administrative, non-partisan issues so that the other two (or more) would resolve to party-specific policy issues. Ideally, the result would then show significant[7] differences in the probabilities of authors being assigned to partisan topics.

Alternatively, the more commonplace LDA is another feasible way of modeling the way in which websites receive a makeover when party control of the mayorship changes. As noted by (Rosen-Zvi et al. 2004), LDA can be considered a special case of the author-topic model, "where each document has one unique author". Here, a city website would be a collection of documents, each of which originates entirely from either Democratic or Republican leadership. In a way, this captures the data-generating process more closely.

Even so, neither the author-topic model nor the LDA model are a perfect fit for the structure of city government websites. One assumes that multiple authors write one document. This may be appropriate for the original purpose of author-topic models, scientific papers, which have multiple authors. This is not as good of a fit for city governments however. Using the LDA instead deals

[6]This may be wrong in some cases, as they might be leftovers from yet another predecessor government. If the WaybackMachine data reaches back far enough, we might be able to identify and include only the documents that were added since the previous election.

[7]Though I am not sure how to calculate that.

with this problem, but it also assumes that one document has multiple topics. For city government websites, this may be an erroneous assumption, because in many cases, each of these files appears to have one very specific purpose - agricultural, parks and recreation, smoke detectors, and so on.

Furthermore, if we only investigate cities in which control of government changes from one party to another, we may overestimate its effect. Not only does a transition in party control occur, but the person in charge also changes. Parties are fairly homogeneous, so that two mayors from the same party may have very different policy preferences and managerial styles. To remedy this problem, we [could] utilize matching, pairing our cases with similar cities in which the incumbent does not run for re-election, but party control stays the same nevertheless.

Another problem with these generative models is that they ignore the fact that despite possible changes to websites due to a leadership transition, large parts of the content carry over. This means that unless the successor government decides to delete everything, some of the existing documents will be preserved, and in the model, also attributed to the new 'author'. But the reverse is not possible, because the predecessor government can't choose to retain documents from the future.

One solution might be to rely on LDA models, but only to apply them to documents that were deleted, the strongest possible expression of a successor governments diverging policy preferences. This way, we could identify directly which issues are considered to be the most divisive.

Traditional author models that rely on lexical or syntactic differences may also be of use. Parties and their voters are associated with specific educational backgrounds, which may in turn be associated with specific writing styles. However, it seems likely that personal rather than partisan authorship is playing an even larger role here.

Another option, that I have not looked closely at yet, are (continuous) dynamic topic models by David Blei, which would allow us to model the progression of topics over time, i.e. when party control changes.

References

Stephan G Grimmelikhuijsen. Transparency of public decision-making: Towards trust in local government? *Policy & Internet*, 2(1):5–35, 2010.

Stephan G Grimmelikhuijsen and Eric W Welch. Developing and testing a theoretical framework for computer-mediated transparency of local governments. *Public administration review*, 72(4): 562–571, 2012.

Ibrahim H Osman, Abdel Latef Anouze, Zahir Irani, Baydaa Al-Ayoubi, Habin Lee, Asım Balcı, Tunç D Medeni, and Vishanth Weerakkody. Cobra framework to evaluate e-government services: A citizen-centric perspective. *Government Information Quarterly*, 31(2):243–256, 2014.

M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494, 2004. ISSN 01689002. doi: 10.1016/j.nima.2010.11.062. URL http://portal.acm.org/citation.cfm?id=1036902.

Lili Wang, Stuart Bretschneider, and Jon Gant. Evaluating web-based e-government services with a citizen-centric approach. In *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on*, pages 129b–129b. Ieee, 2005.

File type	Occurrences
.pdf	1371
.html	819
.png	210
.jpg	131
.gif	99
.js	51
.PDF	50
.aspx	43
.doc	35
.css	32
.JPG	26
.Net	12
.xlsx	6
.docx	5
.ttf	3
.xml	3
.htm	2
.woff	2
.xls	2
.asp	1
.eot	1
.GIF	1
.ico	1
.PNG	1
.ppt	1
.swf	1
.txt	1

Table 1: File types in scraped websites

Website	Files	Size (MB)
brownsvilletn.gov	188	14328
www.centralpointoregon.gov	150	137440
www.dedham-ma.gov	603	212572
www.duncanok.gov	84	47064
www.dutchessny.gov	110	291376
www.ennistx.gov	200	26244
www.greenvillenc.gov	333	25732
www.romi.gov	491	112584
www.trumbull-ct.gov	787	191540
www.westonct.gov	861	213140

Table 2: Test websites

	tf
tax	176324
date	98949
due	97192
amt	96382
town	86726
value	81119
total	70825
parcel	63589
county	56201
market	51758
east	51357
full	51199
nrth	50566
book	50306
deed	50231
bill	49719
acres	48935
acct	44871
csd	43792
owners	43510
res	41803
family	36883
fire	35156
school	33685
name	30382
red	30362
taxable	30248
hook	29902
homestead	29287
outside	28593

Table 3: Top term frequencies for 10 test websites