View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by ScholarWorks@UMass Amherst

University of Massachusetts Amherst
ScholarWorks@UMass Amherst

JITP 2011: The Future of Computational Social Science

The Journal of Information Technology and Politics Annual Conference

2011

InfoExtractor – A Tool for Social Media Data Mining

Chirag Shah
Rutgers University - New Brunswick/Piscataway, chirags@rutgers.edu

Charles File
Rutgers University - New Brunswick/Piscataway, chasfile@rutgers.edu

Follow this and additional works at: https://scholarworks.umass.edu/jitpc2011

Part of the Communication Commons, Computational Engineering Commons, Computer Engineering Commons, Political Science Commons, Public Affairs, Public Policy and Public Administration Commons, and the Science and Technology Studies Commons

Shah, Chirag and File, Charles, "InfoExtractor – A Tool for Social Media Data Mining" (2011). *JITP 2011: The Future of Computational Social Science*. 7.
Retrieved from https://scholarworks.umass.edu/jitpc2011/7

This Article is brought to you for free and open access by the The Journal of Information Technology and Politics Annual Conference at ScholarWorks@UMass Amherst. It has been accepted for inclusion in JITP 2011: The Future of Computational Social Science by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.

First Submission: 01/27/2011

Revised Submission: 05/06/2011

Accepted:

InfoExtractor – A Tool for Social Media Data Mining

Chirag Shah and Charles File

School of Communication and Information

Rutgers, The State University of New Jersey

New Brunswick NJ 08901 USA

chirags@rutgers.edu, chasfile@rutgers.edu

Abstract

We present InfoExtractor, a web-based tool for collecting data and metadata from focused social media content. InfoExtractor then provides this data in various structured and unstructured formats for easy manipulation and analysis. The tool allows social science researchers to easily collect data for quantitative analysis, and is designed to deliver data from popular and influential social media sites in a useful and easy to access way. InfoExtractor was designed to replace traditional means of content aggregation, such as page scraping and brute-force copying.

Keywords

InfoExtractor; social media; politics; YouTube; Wikipedia; data mining; research tools; content analysis.

Introduction

Social media platforms and services such as Facebook, Twitter, YouTube, FourSquare, Flickr, WordPress, and Tumblr make more data available regarding people's lives, intentions, thoughts, activities, and attitudes than ever before. As these services are shifting the social and communication infrastructure of our society, having proper tools and techniques to study these platforms becomes ever more critical for understanding social activities (Naaman, Boase, & Lai 2010), public opinion (Bennett & Iyengar, 2008; Castells, 2009), political action (Parsons, 2010; Siegel, 2009) and more. At the same time, the ubiquity of new social media has stupefied some of the best students of social change, with the prominent exceptions of marketing, advertising, and information science researchers (Blythe & Cairns, 2009; Xenos, 2010). Meanwhile, transformative social and political norms are emerging in online social practices and their off-line corollaries (Glynn, Huge, & Lunney, 2009; Wallsten, 2010). The pace of change and innovation is a significant challenge for researchers armed only with traditional tools and techniques (e.g., spreadsheets). These researchers currently lack the tools to collect and analyze social media at scale.

Online media have a great number of advantages for social science researchers: they provide lasting and exacting accounts of discussions, they contain useful metadata (date, time, subject tags, user names, etc.), they often contain comment discussions, and so on. Unfortunately, this data tends to be difficult to retrieve from many sites. Further compounding the problem, this data is difficult to retrieve in different ways – creating a page scraper for one site doesn't help retrieve data from others, nor does it help when that site's design changes. InfoExtractor was conceived to solve these problems.

In this workbench note, we describe InfoExtractor, a tool to extract the metadata and comments for a variety of online media. It is designed with a number of features that let the users collect and collate their data in useful ways. In addition to extracting data from single pages, InfoExtractor allows for data to be collected from several sites at once using a batch extraction feature. It also allows for data to be collected automatically by third party software applications through an Application Programming Interface (API). Furthermore, InfoExtractor allows users to retrieve their data in a number of different formats, including text, comma separated values (CSV), and Extensible Markup Language (XML). This flexibility of both input and output increases InfoExtractor's usefulness.

InfoExtractor works seamlessly with other projects developed by the primary author in order to facilitate the aggregation of online data. Pages catalogued in ContextMiner[1] – a tool for collecting links to social media that satisfy a number of search criteria – can be quickly analyzed in InfoExtractor. This tool allows for the collection of extended data and metadata related to the pages that ContextMiner collects links to. InfoExtractor was therefore designed to be tool for collecting more in-depth data on a particular page or set of pages, while ContextMiner provides more general summary data on a large number of pages (Shah, 2009).

InfoExtractor was originally designed to mine YouTube data as part of a project to capture and catalogue political communication. Since then, it has been expanded to be useful to social scientists interested in answering any number of research questions. InfoExtractor's key strengths include:

1. Ease of use

[1] http://www.contextminer.org

2. Structured data output in a variety of formats

3. An automated API

InfoExtractor is therefore well-suited to answering initial, investigative queries about social media. Social media research is different from many social science pursuits in that the data is not only abundant, but often cripplingly over-abundant. Many involved in social media research note the "fire hose of data" effect that comes with mining social media data. It can be difficult to make exploratory investigations, and to quickly test social media research hypotheses, without drowning in a sea of data. Many social media researchers, in fact, collect massive amounts of data first, then form research hypotheses, then cull the data they have into something more useful and meaningful. InfoExtractor can help fill this gap in the tool-set that exists between initial research hunches and purposeful, long-term data collection, because it provides real-time results in an easy to analyze format. Researchers can use InfoExtractor to quickly bridge the gap between research question and research design.

InfoExtractor's direct interface does not require sign-up or login. Nor must a user wait for results to be collected and collated, as with many other social media data mining tools. Instead, InfoExtractor presents direct, immediate collection of a wide variety of social media data. Further, this data is easily retrieved in XML, CSV, Text, or HTML formats. These data formats are easily importable into Microsoft Excel, SPSS, R, or any number of other data analysis packages. The quick, simple, and easy-to-use workflow design of InfoExtractor allows researchers to collect a wide variety of data from a wide variety of sources, quickly import them into analysis software, and test initial hypotheses. InfoExtractor can greatly assist researchers seeking to explore and test early research hypotheses.

With its built-in API, InfoExtractor also provides for more long-term data collection. In conjunction with analysis tools that support RSS feed data import – like DiscoverText – researchers can leverage InfoExtractor's capabilities over a period of time automatically. In addition, researchers seeking to implement their own custom social media data mining solutions are saved the odious and time-consuming task of developing page-scraping algorithms. They can simply connect their software to InfoExtractor's API, which will provide the data in a structured format to them. This will allow them to focus their programming resources on more important tasks like data analysis.

Use

There are three main ways to use InfoExtractor to retrieve data from the web pages it supports:

1. Using the text box in order to extract data from a single web page.
2. Using the file upload feature to extract data from multiple web pages.
3. Using InfoExtractor's API to programmatically extract data from a web page.

In order to extract information from a webpage, the URL for the page that the user is interested in is entered into the text box on the InfoExtractor home page. The user then clicks the "InfoExtract" button. On the next page, the user will see links to links to download the extracted data in text, CSV, or XML formats. Note that from here, some pages that InfoExtractor supports – YouTube, for instance – will have the number of comments listed and linked. If the comments are desired, the user can click on that link. The resulting page will return the comments in a simplified format, as well as provide links to download these comments in a structured CSV or XML data format as with the original page data.

In order to extract data from multiple pages using a single text file, The user creates a text file that has each link they would like to extract data from on a separate line. The user then uploads this file to InfoExtractor by clicking the "Choose File" button on the InfoExtractor home page. The user then clicks the "InfoExtract" button. As before, the user will see a link to the data in structured format, though in this case it is only available in XML. (Because different sites that InfoExtractor supports return different data, placing data from different websites in a single text- or CSV-based table doesn't make much sense).

Using InfoExtractor's API to programmatically extract data from a web page. By sending HTTP GET requests to InfoExtractor's API either directly through a web browser or through another software tool) users can automatically extract data programmatically. The API is located at: http://www.infoextractor.org/autoExtract.php

The API supports the options shown in Table 1.

Table 1
InfoExtractor API options.

Option	Use	Value	Effect
url	Required	A hex-encoded URL to a page that InfoExtractor supports.	InfoExtractor returns the data extracted from the page pointed to by the URL in this variable.
format	Optional	'xml' [default] 'csv' 'txt' 'html'	InfoExtractor returns the data extracted from the page pointed to by the "url" variable in the format specified by this variable.
comments	Optional	Any string beginning with 'y' or 'Y' or '1'	InfoExtractor returns, along with the data

	indicates “on.”	extracted from the page pointed to by the “url” variable, the comments and the data associated with these comments from that page as well.

Examples:

- http://www.infoextractor.org/autoExtract.php?url=http://www.youtube.com/watch%3Fv%3DQJBhRbsyrJo
- http://www.infoextractor.org/autoExtract.php?format=csv&url=http://www.youtube.com/watch%3Fv%3DQJBhRbsyrJo
- http://www.infoextractor.org/autoExtract.php?format=txt&comments=yes&url=http://www.youtube.com/watch%3Fv%3DQJBhRbsyrJo

Case Studies

InfoExtractor has a number of potential uses for social scientists in general, and for political scientists in particular. Because an increasing amount of the debate surrounding political discourse is happening online, gathering data about this discourse from online sources is increasingly important. The scale and ferocity of these discussions tends to ramp up when the argument is over facts rather than mere opinion. One example of this phenomenon is Wikipedia. On this site, any user may create or edit an article, and it has leveraged this crowd-sourcing to become one of the most important repositories of knowledge in the world. Given both its openness and scale, Wikipedia has become ground zero for the kind of detail-based factual debates that increasingly dominant contemporary political discourse (e.g., “how and why did John Kerry win his Purple Heart Medal?”: over 12,000 edits; “was conservative commentator

Ann Coulter born in 1961 or 1963?": over 10,000 edits). As such, analyzing the debates that surround the creation of consensus – such as it is – around these facts can provide important and useful insight into the politics of public understanding and opinion.

YouTube provides an interesting counter-point; where Wikipedia is a site ostensibly dedicated to the creation and maintenance of consensus-based truth, YouTube provides a forum for the distribution personal opinions on a massive scale. Popular YouTube videos routinely receive millions of views, placing them on par – in terms of audience size – with television. Importantly, however, YouTube removes almost entirely the cost and access barriers to entry present in television. Increasingly, therefore, politicians are using YouTube to deliver their message directly to voters, removing the editorial and commentary frames that surround their messages on traditional news media. Indeed, some politicians have even taken to making announcements and news releases on YouTube, eschewing the traditional press release model entirely. Collecting and analyzing data from YouTube, then, is increasingly essential to understanding the political activity of an ever-increasing number of politicians.

Case Study 1: New Jersey Governor Chris Christie on YouTube

New Jersey Governor Chris Christie has become something of a YouTube sensation. As of Januray, 2011 he has 182 videos posted to YouTube and has received over 2.3 million views. This is especially remarkable considering that Christie was sworn in in January of 2009, meaning that he is averaging a video posting to YouTube more than once every other day. Clearly, his YouTube messaging is an important part of Christie's overall media plan. Indeed, a *New York Times* article that calls Christie a "YouTube star" points out that his YouTube presence "dwarfs those of his peers around the country and has fueled the buzz about his being a potential national candidate" (Pérez-Peña, 2010).

Christie is famous for his forthright speaking manner and making direct connections with voters. He holds a large number of "town hall" meetings and question-and-answer sessions with voters, many of which are filmed and placed on YouTube. By using InfoExtractor, we can see what types of videos he posts, what type of content they contain, and what people are saying about them.

First, as demonstrated in Figure 1, we can use InfoExtractor to quickly retrieve information about Gov. Christie's YouTube account[2].

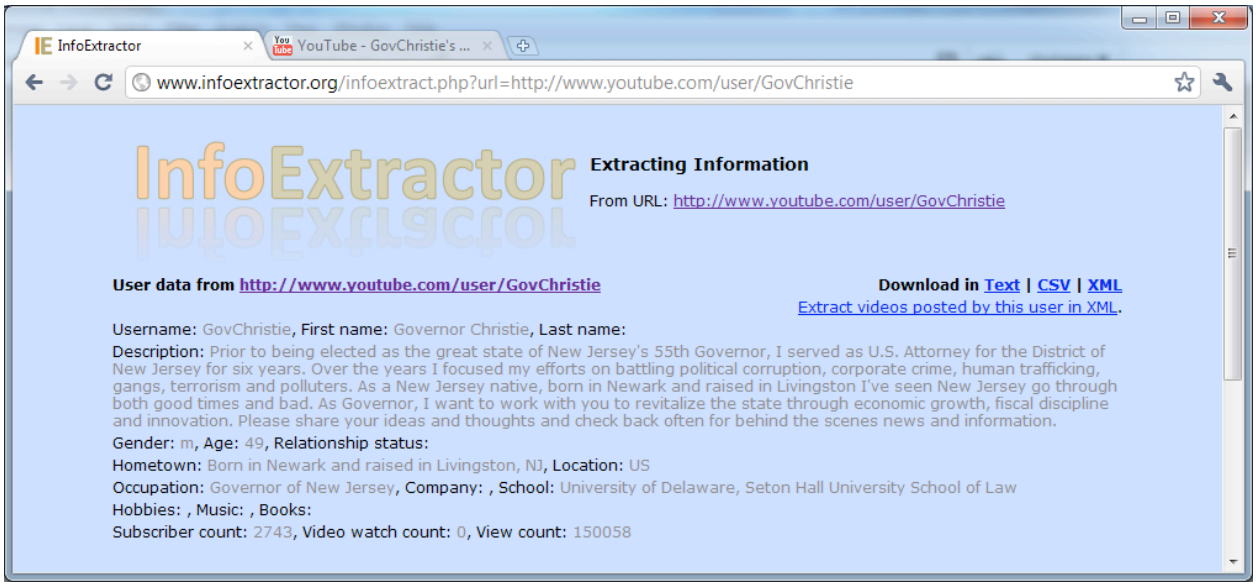
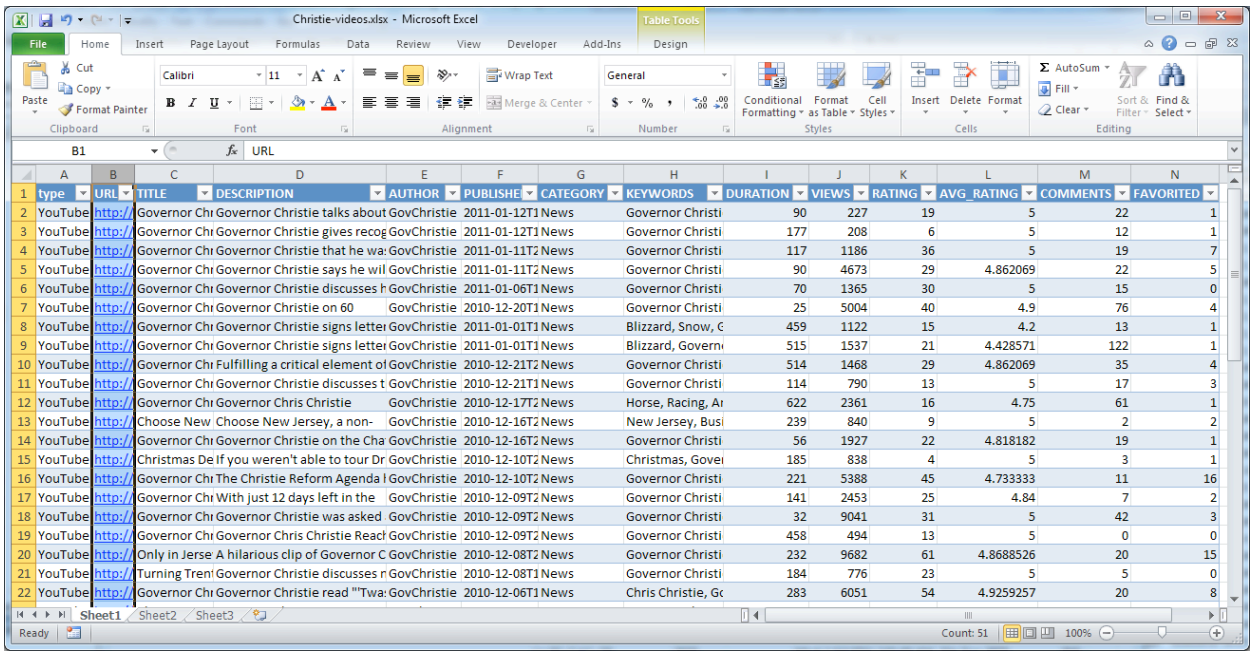


Figure 1. InfoExtractor results page for YouTube account of Gov. Chris Christie.

By using the "Extract videos posted by this user in XML" link, we can easily extract the data for each of Gov. Christie's videos. Figure 2 shows the results of that extraction.

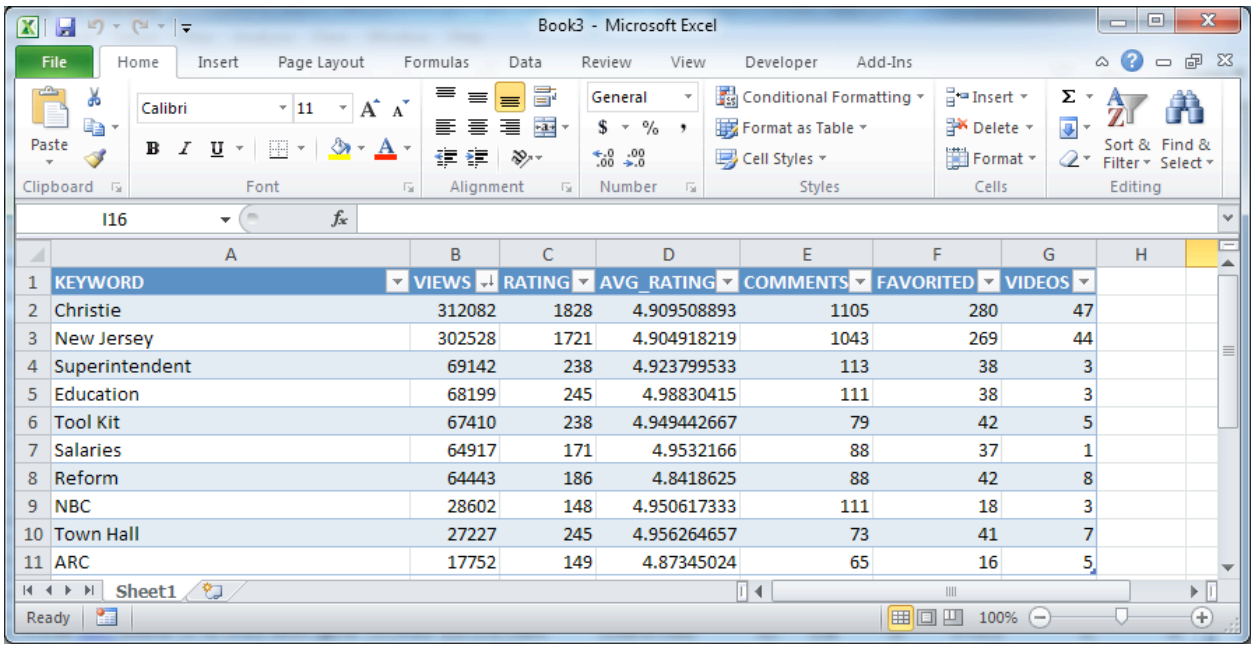
[2] Note that the primary author of this paper also developed TubeKit (http://www.tubekit.org) for creating customized YouTube harvesters (Shah, 2010). In contrast, InfoExtractor allows users to collect metadata and other social media data of a YouTube video or channel using a web-based interface.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	type	URL	TITLE	DESCRIPTION	AUTHOR	PUBLISHE	CATEGORY	KEYWORDS	DURATION	VIEWS	RATING	AVG_RATING	COMMENTS	FAVORITED
2	YouTube	http://	Governor Chr	Governor Christie talks about	GovChristie	2011-01-12T1	News	Governor Christi	90	227	19	5	22	1
3	YouTube	http://	Governor Chr	Governor Christie gives recog	GovChristie	2011-01-12T1	News	Governor Christi	177	208	6	5	12	1
4	YouTube	http://	Governor Chr	Governor Christie that he wa	GovChristie	2011-01-11T2	News	Governor Christi	117	1186	36	5	19	7
5	YouTube	http://	Governor Chr	Governor Christie says he wil	GovChristie	2011-01-11T2	News	Governor Christi	90	4673	29	4.862069	22	5
6	YouTube	http://	Governor Chr	Governor Christie discusses h	GovChristie	2011-01-06T1	News	Governor Christi	70	1365	30	5	15	0
7	YouTube	http://	Governor Chr	Governor Christie on 60	GovChristie	2010-12-20T1	News	Governor Christi	25	5004	40	4.9	76	4
8	YouTube	http://	Governor Chr	Governor Christie signs letter	GovChristie	2011-01-01T1	News	Blizzard, Snow, C	459	1122	15	4.2	13	1
9	YouTube	http://	Governor Chr	Governor Christie signs letter	GovChristie	2011-01-01T1	News	Blizzard, Governo	515	1537	21	4.428571	122	1
10	YouTube	http://	Governor Chr	Fulfilling a critical element of	GovChristie	2010-12-21T2	News	Governor Christi	514	1468	29	4.862069	35	4
11	YouTube	http://	Governor Chr	Governor Christie discusses t	GovChristie	2010-12-21T1	News	Governor Christi	114	790	13	5	17	3
12	YouTube	http://	Governor Chr	Governor Chris Christie	GovChristie	2010-12-17T2	News	Horse, Racing, Ar	622	2361	16	4.75	61	1
13	YouTube	http://	Choose New	Choose New Jersey, a non-	GovChristie	2010-12-16T2	News	New Jersey, Busi	239	840	9	5	2	2
14	YouTube	http://	Governor Chr	Governor Christie on the Cha	GovChristie	2010-12-16T2	News	Governor Christi	56	1927	22	4.818182	19	1
15	YouTube	http://	Christmas De	If you weren't able to tour Dr	GovChristie	2010-12-10T2	News	Christmas, Gover	185	838	4	5	3	1
16	YouTube	http://	Governor Chr	The Christie Reform Agenda I	GovChristie	2010-12-10T2	News	Governor Christi	221	5388	45	4.733333	11	16
17	YouTube	http://	Governor Chr	With just 12 days left in the	GovChristie	2010-12-09T2	News	Governor Christi	141	2453	25	4.84	7	2
18	YouTube	http://	Governor Chr	Governor Christie was asked	GovChristie	2010-12-09T2	News	Governor Christi	32	9041	31	5	42	3
19	YouTube	http://	Governor Chr	Governor Chris Christie Reach	GovChristie	2010-12-09T2	News	Governor Christi	458	494	13	5	0	0
20	YouTube	http://	Only in Jerse	A hilarious clip of Governor C	GovChristie	2010-12-08T2	News	Governor Christi	232	9682	61	4.8688526	20	15
21	YouTube	http://	Turning Tren	Governor Christie discusses r	GovChristie	2010-12-08T1	News	Governor Christi	184	776	23	5	5	0
22	YouTube	http://	Governor Chr	Governor Christie read "Twa	GovChristie	2010-12-06T1	News	Chris Christie, Gc	283	6051	54	4.9259257	20	8

Figure 2. InfoExtractor data for YouTube videos posted by Gov. Chris Christie in Excel.

Because InfoExtractor provides this data to the user in a structured format, data analysis proceeds easily. For instance, it might be interesting to look at the keywords Christie most often uses to tag his videos. Figure 3 shows the top ten such keywords, arranged by the total view count for the videos to which those keywords were applied.



	A	B	C	D	E	F	G	H
1	KEYWORD	VIEWS	RATING	AVG_RATING	COMMENTS	FAVORITED	VIDEOS	
2	Christie	312082	1828	4.909508893	1105	280	47	
3	New Jersey	302528	1721	4.904918219	1043	269	44	
4	Superintendent	69142	238	4.923799533	113	38	3	
5	Education	68199	245	4.98830415	111	38	3	
6	Tool Kit	67410	238	4.949442667	79	42	5	
7	Salaries	64917	171	4.9532166	88	37	1	
8	Reform	64443	186	4.8418625	88	42	8	
9	NBC	28602	148	4.950617333	111	18	3	
10	Town Hall	27227	245	4.956264657	73	41	7	
11	ARC	17752	149	4.87345024	65	16	5	

Figure 3. Top keywords (by views) of videos posted by Gov. Chris Christie on YouTube.

From this quite cursory analysis, we can quickly see that a large number of the views of Christie's videos are of those videos dealing with education issues: aside from the de riguer "Christie" and "New Jersey" tags, the top two tags deal with education. Indeed, though other keywords were applied to a larger number of videos, the videos that were tagged with education-related keywords, though fewer in number, received far more views. In fact, the New York Times profile mentioned above points out that Christie's most viewed video involves a verbal sparring match he has with a teacher over his education policies (Pérez-Peña, 2010). It might be interesting to study how Christie's YouTube notoriety – and the subsequent attention in traditional media that this notoriety engendered – based on this issue has shaped his policy agenda. In this way we can see how InfoExtractor can guide researchers

As this cursory study has demonstrated, Christie is most well-known online for his statements on education, which may seem odd given his background as a former attorney

general. One wonders whether Christie's agenda is driving his YouTube fame, or if his somewhat surprising YouTube fame has begun driving his agenda. Using InfoExtractor and some basic analytical techniques, it is quick and easy to discover trends and identify fruitful avenues for research in social media like these.

Case Study 2: Supreme Court Justice Clarence Thomas on Wikipedia

United States Supreme Court Justice Clarence Thomas is widely regarded to be an enigma. His reticent, reserved nature is almost widely noted. Despite holding a position of incredible power, Thomas is notoriously reluctant to discuss his opinions or policies in either a personal or a professional setting. A 2009 *New York Times* article noted that he had not asked a question during oral arguments before the court in over four years, and he rarely writes majority opinions (Liptak, 2009). Though he assumed his position on the bench in 1991, it was not until 2007 that Thomas appeared in his first television interview, a *60 Minutes* piece entitled "Clarence Thomas: The Justice nobody knows" (Radutzky & Cetta, 2007).

Yet despite being – compared to many of his colleagues on the court – almost totally invisible in both a professional and personal sense, discussion of Thomas rages in heated fashion on Wikipedia. His personal biography page has more discussion and has been edited more times than that of any other sitting Supreme Court Justice. It seems bizarre that the "Justice nobody knows" should also be the same one that everyone on Wikipedia wants to talk about. Using InfoExtractor, we can learn more about what drives this seeming contradiction between the silent Justice and the vociferous discussion about him.

Using InfoExtractor for Wikipedia discussion proceeds in a similar manner as it does for YouTube. After copying the link for Clarence Thomas's Wikipedia page into the InfoExtractor

search box and clicking the “InfoExtract” button, the user is presented with a number of options for downloading the data. In the Figure 4 the structured nature of this data – specifically in the XML format – is demonstrated.

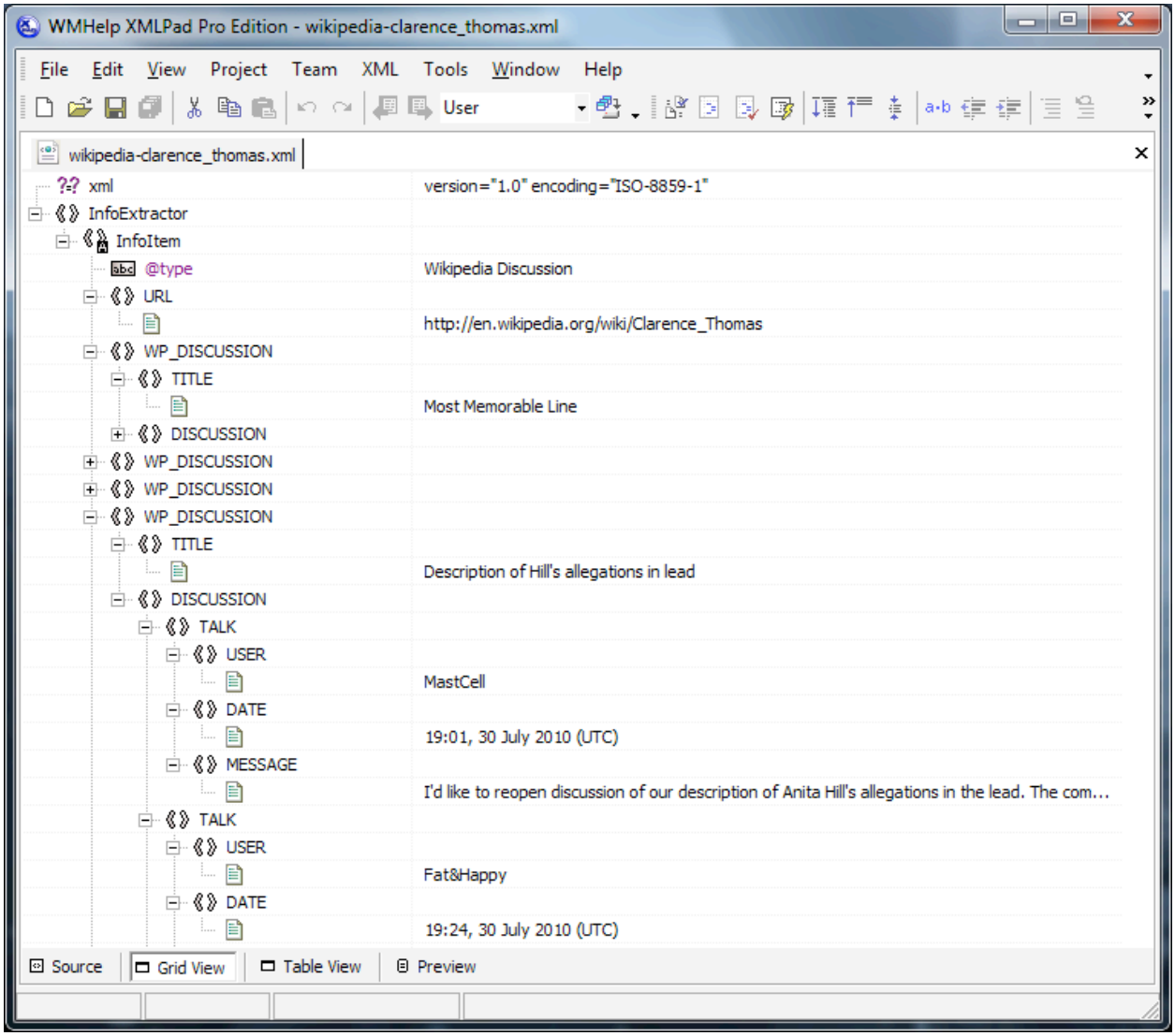


Figure 4. The XML structure of data on the discussion page for Clarence Thomas on Wikipedia.

Here some of the behind-the-scenes features of InfoExtractor are demonstrated. Even though Wikipedia does not provide access to its raw data, InfoExtractor is nonetheless able to mine this data from the site while maintaining its structure and context. In the example above, we

can see how individual user posts are organized into nodes, containing the user and date associated with each post. These nodes are then organized into threaded discussions, which each post node situated directly below the one it is a reply to. Further, these discussion thread nodes are grouped as the children of a topic node, so that discussions of a single topic are grouped together. This discussion node also has metadata associated with it that is mined from the Wikipedia page. Finally, these topic nodes are all children of the page node, which itself contains more metadata about the Wikipedia page.

This structuring of the data has several advantages. One such advantage is that by collecting both structure and metadata (in addition to the comments themselves) we are able to analyze and visualize this data in useful ways. For example, Figure 5 is a translation of the XML data in Figure 4 into a timeline, showing the time, topic, and number of posts to Clarence Thomas's Wikipedia discussion page.

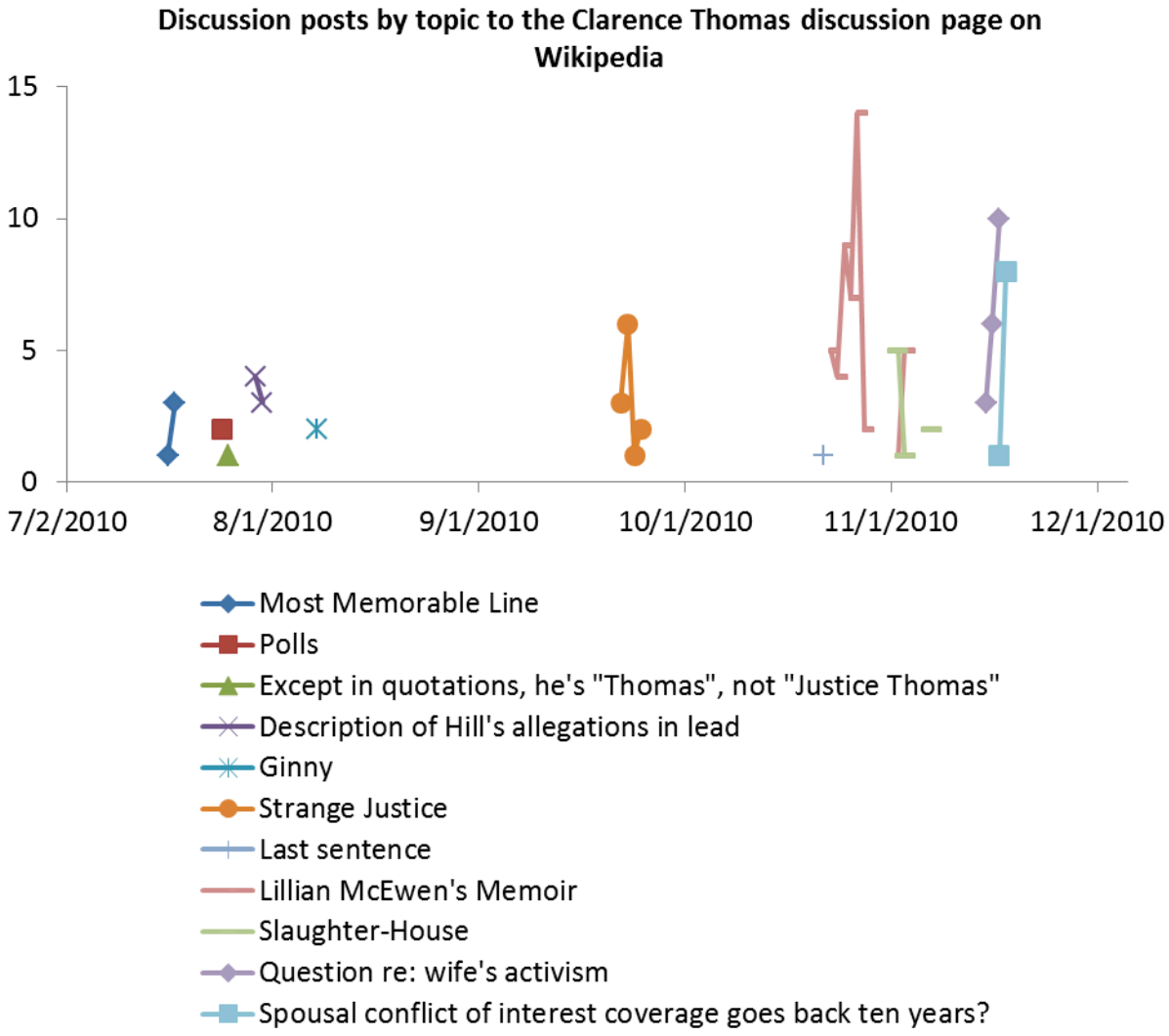


Figure 5. Discussion posts by topic to the Clarence Thomas discussion page on Wikipedia.

This chart plots the number of comments posted to each of the various topics in the discussion page of Clarence Thomas's Wikipedia entry over a period of several months in late 2010. Though this chart includes the type of raw frequencey distribution that merely counting the entries would provide, it also contains aditional information that is only possible to express because InfoExtractor provides rich metadata in its XML data.

This is a brief, illustrative example of an analysis of Wikipedia comments. However, several features of the discussion are prevalent, and these features might be used to help direct more rigorous research. The first and perhaps most obvious feature of this chart is its punctuated equilibrium: rather than a continuous, steady stream of discussion and changes to the article, it appears that the dialogue is characterized by brief flurries of activity, and then is silent, sometimes for weeks at a time. The second prminent feature is the strong topic-driven nature of the discussion. Topics seem to come up, be discussed, and then settle down rather quickly. Rather than long, drawn out discussions and flame wars, it seems that some sort of decision or consensus is able to be reached among users in a matter of days. This, especially, seems remarkable. In a political climate characterized by unprecedented partisan vitriol, the data provided by InfoExtractor seem to demonstrate the surprising capacity a forum like Wikipedia has for facilitating consensus and ameliorating conflict. While much of the media becomes ever more partisan, Wikipedia seems to promote rapid and effective consensus-building even over some of history's most torrid scandals (Anita Hill's allegations of sexual harassment of her by Thomas, in this example for instance). Clearly, how this collaboration process takes place and what about Wikipedia makes it work so well as a political consensus-builder is a rich area for further study.

Using InfoExtractor Data

Because of its flexible, structured format, data from InfoExtractor can be easily accessed and applied to a wide variety of uses. InfoExtractor also extracts a great deal of metadata information, allowing for a large number of possible applications. Several are discussed below.

Structured Data Output

InfoExtractor allows for the export of structured data containing a variety of data, in addition to the in-depth examples given above. For instance, it provides the data for comments on YouTube videos, which can provide interesting insight into the types of discussions that occur around a video. For instance, by extracting the comments from Gov. Christie's most popular video, we can discover the nature of the voices involved in the discussion. The data was exported as a CSV file into Microsoft Excel, and then a frequency distribution of the number of posts by the varying comment authors was prepared. The plot of this distribution demonstrates that a relatively small number of users are driving the discussion: the top 12% most frequent commenters made as many comments as the bottom 88%.

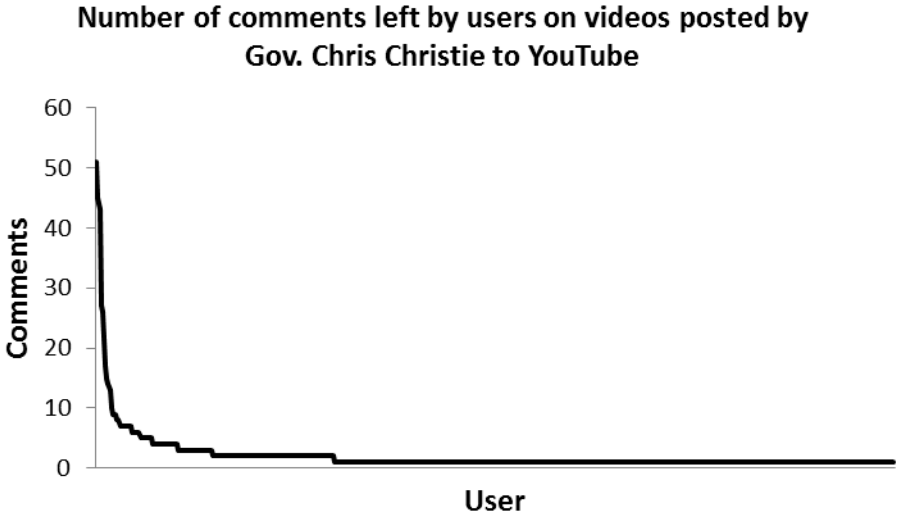


Figure 6. Number of comments left by users on videos posted by Gov. Chris Christie to YouTube.

Text for Content Analysis

InfoExtractor also allows data to be downloaded in simple text format. This allows for easy import of the data into any of a number of content analysis software packages. To illustrate, again using InfoExtractor, the content of all the comments left on Gov. Christie's most popular video was analyzed in Yoshikoder[3], an open source content analysis software package. A portion of the simple word frequency analysis which resulted is shown in Figure 7.

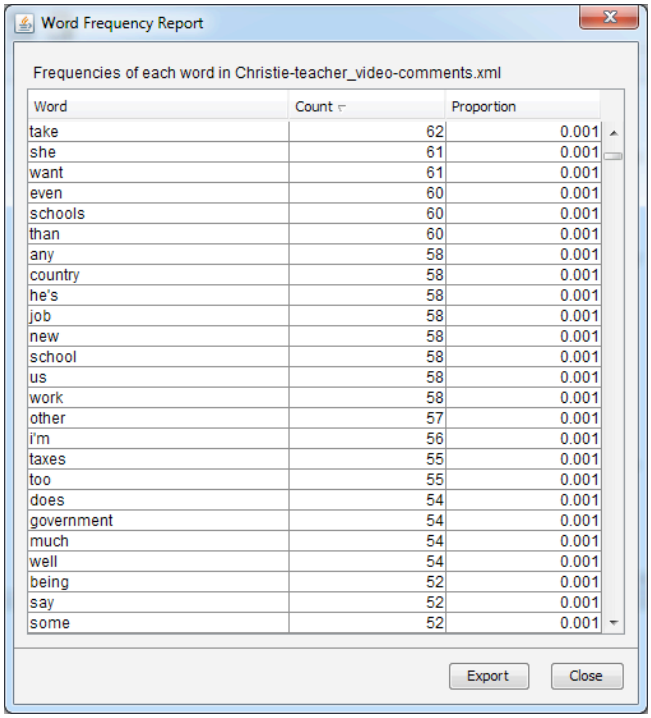


Figure 7. Word frequency analysis using Yoshikoder of comments to Gov. Chris Christie's most popular video on YouTube.

[3] http://www.yoshikoder.org/

API Access

InfoExtractor allows for the automatic culling of data into a third-party data analysis tool using its API and XML. For instance, the web-based content analysis tool DiscoverText[4] allows the import of live XML data over an RSS feed. Using this feature, it is possible to easily import data gathered from InfoExtractor into DiscoverText. This type of integration has several advantages, as it allows for direct connection between the data and the analysis tool, eliminating the need for downloads, and also allows the analysis to be constantly updated as the data grows and changes.

Limitations

InfoExtractor, like all tools, is not without its limitations. It is designed to work as a solution to certain problems: the initial collection of social media data, and the long-term collection of data in conjunction with other existing or custom tools. InfoExtractor is not an end-to-end social media research data collection and analysis solution.

InfoExtractor is designed to provide on-demand, real-time collection of social media data. It does not have the capabilities to store data, nor can it collect historical data, other than what is currently in the datasets of the sites it mines. InfoExtractor also does not provide analysis tools, it only collects the data and delivers it in a structured format. This data can be imported into a third-party data analysis package, but cannot be analyzed within InfoExtractor. Finally, InfoExtractor is limited in the number of sources from which it can mine social media data. Unfortunately, there is no set standard for social media collection, and many social media services do not even provide their data for easy collection through an API. These limitations are

[4] http://www.discovertext.com/

what necessitated the creation of InfoExtractor in the first place, and they continue to represent roadblocks to social media research. Through its API, InfoExtractor can provide a solution to other social media research tool developers seeking to circumvent these roadblocks. They can focus on other, more interesting aspects of tool development, and access the data collection capabilities of InfoExtractor through its API.

Conclusion

InfoExtractor was designed to permit social science researchers to retrieve data and metadata in structured formats from popular online media sites. It was designed in such a way that it would not be dependent on page scraping, and would free users from having to perform such scraping and copying themselves. It is designed to be maximally flexible in the data formats it provides, and as such is well-suited to any number of quantitative research approaches and methods. InfoExtractor provides important capabilities to social science researchers interested in studying online media. It can provide structured data that can be easily manipulated and analyzed.

As online media become increasingly important in all aspects of society, the study of them becomes a concomitant research imperative. However, the technical barriers that confront many traditional social scientists often confound the effective pursuit of these important areas of research. It is our hope that InfoExtractor can serve to remove some of these barriers, and that it may provide researchers a valuable tool for the conduct of their important work.

References

Bennett, W. L., & Iyengar, S. (2008). A new era of minimal effects? The changing foundations of political communication. *Journal of Communication*, *58*(4).

Blythe, M., & Cairns, P. (2009). Critical methods and user generated content: The iPhone on YouTube. *Proceedings of the 27th International Conference on Human Factors in Computing Systems* (pp. 1467-1476).

Castells, M. (2008). The new public sphere: Global civil society, communication networks, and global governance. *The Annals of the American Academy of Political and Social Science*, (616), 78-93.

Glynn, C. J., Lunney, C. A., & Huge, M. E. (2009). The polls - trends: public perceptions of the US residential housing market before, during, and after the housing bubble (1990-2009). *Public Opinion Quarterly*, *73*(4), 807-832.

Liptak, A. (April 14, 2009). Reticent justice opens up to a group of students. *New York Times*, A11.

Naaman, M., Boase, J., & Lai, C.-H. (2010). Is it really about me? Message content in social awareness streams. *CSCW 2010: Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work* (pp. 189-192). New York, NY: ACM.

Parsons, B. (2010). Social networks and the affective impact of political disagreement. *Political Behavior*, *32*(2).

Pérez-Peña, R. (December 1, 2010). Talking tough and drawing viewers, Christie is a YouTube star. *New York Times*, A26.

Radutzky, M., & Cetta, D. (Producers) (September 30, 2007). Clarence Thomas: The Justice nobody knows. *60 Minutes*. New York, NY: CBS.

Shah, C. (2009). ContextMiner: Supporting the mining of contextual information for ephemeral digital video preservation. *International Journal of Digital Curation*, *4*(2).

Shah, C. (2010). Supporting research data collection from YouTube with TubeKit. *Journal of Information Technology and Politics*, *7*(2/3), 226-240.

Siegel, D. A. (2009). Social networks and collective action. *American Journal of Political Science*, *53*(1), 122-138.

Wallsten, K. (2010). "Yes we can": How online viewership, blog discussion, campaign statements, and mainstream media coverage produced a viral video phenomenon. *Journal of Information Technology and Politics*, *7*(2/3), 163-181.

Xenos, M. (2010). Guest editor's introduction. *Journal of Information Technology and Politics*, *7*(2/3), 89-92.

Author Note

Chirag Shah

School of Communication and Information, Rutgers University

Chirag Shah is an Assistant Professor in the School of Communication and Information at Rutgers University. He received his PhD in Information and Library Science from the University of North Carolina (UNC) at Chapel Hill. He received his MS in Computer Science from UMass Amherst, where he worked with Bruce Croft and James Allan on high accuracy retrieval, and topic detection and tracking. At UNC, he worked with Gary Marchionini and Diane Kelly on various issues concerning exploratory information seeking and interactive information retrieval. He has also worked at many world-renowned research laboratories, such as FXPAL in California and National Institute of Informatics in Tokyo, Japan. His dissertation is focused on collaborative information seeking. He is also interested in social search and question-answering, digital preservation, and contextual information extraction. In addition to InfoExtractor, he has developed several tools for exploratory information seeking and extraction, including "Coagmento" for collaborative information seeking and the award-winning "ContextMiner" for capturing contextual information from multiple online sources.

Correspondence concerning this article should be addressed to Chirag Shah, Rutgers University, 4 Huntington St. New Brunswick, NJ 08901 or to chirags@rutgers.edu.

Charles File

School of Communication and Information, Rutgers University

Charles File is a PhD student in the School of Communication and Information at Rutgers University. He received an MA in Media, Culture, and Communication from NYU in 2008, and a BA Summa Cum Laude in Media and Cultural Studies from Cornell University in 2004.