SOCRATES 2.0: Bridging the Gap between Researchers and Social Media Data through Natural Language Interactions

Dongho Choi, Ziad Matni, Chirags Shah
School of Communication & Information (SC&I)
Rutgers, the State University of New Jersey
4 Huntington St, New Brunswick, NJ 08901
{dongho.j.choi, zmatni, chirags}@rutgers.edu

ABSTRACT
In order to help researchers who have to resort to time-consuming work on collecting, analyzing, and visualizing social media data, we have developed a system, called the Social and Crowdsourced Activity Extraction System (SOCRATES). This tool supports researchers utilize a single, modularly architected, open-sourced, and standardized platform for collecting, analyzing, and visualizing social media data. In this demonstration, a follow-up to one we presented at ASIST 2014 (Matni, Choi, & Shah, 2014), we present “SOCRATES 2.0” with improved functionalities, more precisely, those that help users focus on their research problem(s) without spending time and effort dealing with details in data processing. Through a process of asking questions related to the users’ interests and requests for certain analyses, the system generates one sentence about a possible research statement and provides the collected data and analysis results with one click.

Keywords
Social media, big data, social computation, system, demo.

INTRODUCTION
An increasing number of diverse social media platforms and services, such as Facebook, Twitter, YouTube, and Flickr, make more data available about peoples’ lives, thoughts, activities, intentions, and sentiments than ever before. The shifting of social and communication infrastructure of society places a demand on the proper tools and techniques to study particular platforms for investigating social activities (Kooti, Yang, Cha, Gummadi, & Mason, 2012; Naaman, Boase, & Lai, 2010; Shah, 2010), political action (Parsons, 2009; Siegel, 2009), public opinion (Bennett & Iyengar, 2008; Castells, 2008), and more. Furthermore, the dramatically increasing number of social media keeps researchers challenged.

In attempting to address this problem, the authors have developed the SOCRATES platform (Matni, Choi, & Shah, 2014) to help researchers collect, analyze and explore data obtained from social media sources without having to resort to using external (and often very costly) programming, writing codes, and building up applications. The framework was used in a user study to understand people’s different information searching behavior for different types of sources, such as Web vs. social media (Choi, Matni, & Shah, 2015). It was found that, through the platform, people who have an open-end question find a smaller quantity and a less diverse set of documents from social media than what they might discover when using Web search engines.

SOCRATES 2.0
Improved Functionalities
The 3 main components of SOCRATES are Collect, Analyze, and Explore. The Collect module allows users to gather data from social media, while the Analysis module helps users do some basic analysis, such as statistical analysis and textual analysis. The Explore module provides users with ways to visualize the collected data and the results of analysis. In an effort to keep up with the latest social media services’ application programming interfaces (APIs), SOCRATES 2.0 has upgraded certain back-end modules (YouTube API version 3.0 and Facebook API 2.3) and continues to add more data sources.

To provide the usability of stored data throughout the system, we give the users’ an option to download their collected datasets from SOCRATES and also allow them to upload custom datasets of their own from. Additionally, the results of user analysis, in the form of several graphs and images, are available to be downloaded in SVG format, an XML-based vector image supporting interactivity.

Sentence Generation
Using three modules, SOCRATES provides an easier workflow that helps people collect, analyze, and visualize the social media data of their interest. However, we saw several opportunities for improvement on the original version of this tool, especially given those potential users...
who might not have backgrounds and/or practical experience with numerical data with relatively high numbers of variables at a large scale, or in some specific areas of research, such as natural language processing. In order to bridge the gap between those users and the barriers they faced, we developed a feature that helps them think about their research problem(s) before jumping into the modules and picking sources and analysis methods. Instead of working with each component independently and dealing with the details related to the specific features of that component, SOCRATES 2.0 first asks the user several questions to understand his or her research interest(s) and provide results with collected data and the appropriate analysis in at one try.

**SENTENCE GENERATION**
SOCRATES 2.0 has added a layer for sentence generation on top of the three components (Figure 1). The main purpose of this layer is to provide users with a way that helps them define their research interest(s) in one sentence and presents the data and results of relevant analysis at a glance without dealing with data processing details, such as variables, factors, statistics, methods, and so on.

![Figure 1. Natural Language Interaction Layer on top of Main Components](image)

Sentence generation is performed after a set of questions are asked of the users on their research topic, such as what aspects of the topic did they want to investigate, and what type of social media might they be interested in, if any (Figure 2). Upon completing this task, the users are presented with a sentence of research interest and the suggested collected data and analysis result(s) are provided automatically. In the current version of SOCRATES, this is presented as a wizard-based process, with limited options, where the users are walked step-by-step in formulating their search for data and analysis. In future versions of SOCRATES, we intend to provide a natural language interface through which one could express their research questions or hypotheses and then have the system translate that into a workflow involving social media data collection and analysis. Additionally, future versions will have more data collection, analysis, and exploration modules.

![Figure 2. Questions and Workflow of Sentence Generation in the Natural Language Interaction Layer](image)

**EXAMPLE OF THE USE OF THE NATURAL LANGUAGE INTERACTION FEATURE**
Consider a researcher who wants to examine and make sense of the data on social media chatter regarding the International Federation of Association Football (FIFA) scandal. This researcher may not know enough about social media technology to delve into using various APIs and writing computer-programming code, but he is interested in finding out how Twitter users, on the whole, feel about the scandal. Upon starting up the SOCRATES interface, he is asked, “What topic are you interested in?” (Figure 3) and he answers with the acronym “FIFA”. He is then directed to a second question, “Where would you like to get the data from?” along with a choice of all social media modules currently available on the system (Figure 4). Once the researcher clicks on “Twitter” as a choice, he is directed to a third question. In future version of SOCRATES, users will be able to click on multiple sources of social media data. The third question then asks the user, “What would you like to do with the data you collect about these topics?” (Figure 5) and is presented with a choice of all analysis modules currently available. Upon making this last choice, the user is presented with his choices and is given a choice to view the collected data, view the analysis, or go back in the process and engage in different choices (Figure 6). Figure 6 also illustrates what the user would see if he chose to view the collected data, or the analysis results.

**ANOTHER EXAMPLE OF USE**
Another demonstration we can offer is one that directly interfaces with the three main modules of SOCRATES, bypassing the use of the natural language interaction layer. Let us consider that we want to test a hypothesis with data from [reddit](http://www.reddit.com). The hypothesis states that there exists a positive correlative relationship between a [reddit](http://www.reddit.com) post’s comment length and the net votes the comment gets. This can be done with the following steps on SOCRATES:

a. Using the **Collect** component, acquire [reddit](http://www.reddit.com) data using a keyword search.

b. Using the **Analyze** component, create a new variable called “comm_length” which equals the
word count of the “content” variable (supplied by the reddit API).

c. Still using the Analyze component, create a new variable called “net_votes” which equals “upvotes” – “downvotes” (both of which are provided by the reddit API).

d. Still using the Analyze component, run the correlation function between “comm_length” and “net_votes”. Examine the results.

e. Using the Explore component, graph a linear regression graph using the aforementioned two variables, superimposing a scatterplot on an x-y axis.
DEMONSTRATION SETUP
To demonstrate SOCRATES at the conference, we will use a laptop computer connected via the Internet to a server hosting SOCRATES. Visitors will have a chance to try the system out at the demonstration and via the Internet at http://peopleanalytics.org/socrates/.

CONCLUSION
In this paper we presented SOCRATES 2.0, a social-computational system and platform for the study of social media. We have also described two possible scenarios of use. We propose to show this tool’s latest version at the conference, as SOCRATES 2.0.

The demonstration we propose to show at the conference is still a prototype of SOCRATES. In its final form, SOCRATES will also enable intuitive ways to incorporate crowdsourcing functions in all three components of collection, analysis, and exploration of data. SOCRATES will also allow easy integration of data collected outside its system so that it may be used with other modules.

The software architecture of SOCRATES and its data structure implementations are publicly available (https://github.com/InfoSeeking/Socrates). The current version is available for non-commercial usage from http://peopleanalytics.org/socrates/ under a Creative Commons License.

ACKNOWLEDGMENTS
The work reported here is supported by the National Science Foundation (NSF) BCC-SBE award no. 1244704.

REFERENCES


Kooti, F., Yang, H., Cha, M., Gummadi, P., & Mason, W. (2012). The Emergence of Conventions in Online Social Networks. ICWSM.


