Accelerating Social Science Analysis for a New Age

Moving from Traditional Methods for Analyzing Large Scale Text-Based Data to Socially Intelligent High-Performance Computational Methods

Derrick L. Cogburn
American University
4400 Massachusetts Ave NW
Washington, DC 20016
dcogburn@american.edu

Mary E. Hansen
American University
4400 Massachusetts Ave NW
Washington, DC 20016
mhansen@american.edu

Amy Wozniak
American University
4400 Massachusetts Ave NW
Washington, D.C. 20016
aw8813a@american.edu

ABSTRACT (150 WORDS)
The volume of digital data available to social scientists today is increasing exponentially due to the Internet and World Wide Web. Accelerating Social Science for a New Age (ASSANA) is a research project designed to develop, test, refine and disseminate a repeatable interdisciplinary methodology for the computer-assisted analysis of large-scale text-based social science data. We will compare the methods of traditional hand-coding, CAQDAS software, computer assisted keyword and content analysis, and finally sentiment analysis and machine learning to determine the differences between each method; particularly for research questions answered, time, and consistency of findings. Our findings will allow researchers to make better use of the available digital data and to accelerate scientific advancement through engagement of large-N collaborative and comparative studies. It will contribute to HCI by providing a more efficient and effective methodology for data intensive collaboration.

Author Keywords
Text-based data, CAQDAS, sentiment analysis, machine learning, social science, data mining, methodology

ACM Classification Keywords
I.3.6 Methodology and Techniques, miscellaneous

General Terms
Human Factors; Design; Measurement; Performance; Reliability

INTRODUCTION
Because of the growth in use of the Internet and World Wide Web over the last decade by governments, international organizations, businesses, and civil society, the volume of data available to social scientists has increased exponentially. As Berman and Brady [1] note, many social scientists recognize the promise of using large-scale digital data to answer important research questions but do not have the skills or infrastructure to capitalize on them. Many social scientists trained only in traditional methods of qualitative analysis, and even those using Computer Assisted Qualitative Data Analysis Software (CAQDAS) tools, find it difficult to cope with the time and effort required to analyze these large data sources [1, 2]. As a result, many work alone on limited datasets and risk making Type I errors (rejecting a null hypothesis that is true) or Type II errors (accepting a null that is false). Others make extremely slow progress using traditional hand-coded forms of analysis. Moreover, many still work in small, discipline-specific groups that do not incorporate insights from disciplines such as computer science and mathematics.

CSCW'12, February 11–15, 2012, Seattle, Washington, USA.

PROJECT
Accelerating Social Science for a New Age (ASSANA) is a multi-tiered project to develop, refine, and disseminate a methodology for social scientists to use computer-intensive software to engage large-N studies. Four methodologies will be used and compared between three ongoing projects in order to determine the differences in time, consistency in findings, and research questions answered. The four methodologies are described below, from least computer-intensive to most computer-intensive.

Methodologies

1. Traditional Approach: Traditional hand-coding and content analysis where a researcher reads each observation of source material, reflects, makes notes, and develops subject codes along the way.
2. CAQDAS: Computer Assisted Qualitative Data Analysis Software. Enables faster hand-coding and automatic data cleanup. Examples include: Nvivo, QDA Miner, and Atlas.ti.
3. CADM: Computer Assisted Data-Mining and Content Analysis software. Statistical and algorithmic analysis of text to look for patterns. Examples include: WordStat, Nvivo9, SPSS.

4. Sentiment Analysis and Machine Learning: Linguistic algorithmic parsing and analysis of text to uncover potential sentiment. Examples include: Profiler Plus and r/tm.

Text-based Research Projects

We have multiple social science projects that are currently being researched using large-scale text-based data analysis. Pawns to Partners (P2P) has been underway since 2006 and is being used to develop the ASSANA Methodology, while two new projects in Public Diplomacy and Economic History will be used to test and refine this methodology.

1) Pawns to Partners (P2P): This project aims to measure the impact of culture and geographic location on decision-making and leadership in transnational NGOs. We explore the impact of transnational policy-actor networks and epistemic communities on perceived policy power in the United Nations World Summit on the Information Society (WSIS). Specifically, it considers the extent to which membership in an epistemic community, levels of information and communication technology (ICT) policy activity, and perception of the influence of one's own global policy network, help explain perceptions of policy power at WSIS. The datasets for this project are multi-year email archives of transnational advocacy networks organized to facilitate civil society participation in WSIS. The dataset is 121MB and contains 9,942 E-mails.

2) Public Diplomacy: Recently scholars of global governance have asked questions about the role of the Secretary of State in projecting soft power, and explored the differences in scope, focus, and tone, based on delivery location and audience (e.g. international vs. domestic audiences; and international organizations v. countries). At the same time, the US Government has engaged in a very active effort at using social media and other web-based tools to enhance transparency and openness in its primary agency for foreign policy, the US Department of State. Most scholars of international communication refer to this as "public diplomacy" [3, 4, 5] and this research domain is built upon earlier scholarship on "propaganda" and political persuasion [6]. These and other similar efforts across government to enhance transparency and participation, such as the Open Government Initiative (http://www.whitehouse.gov/open), are yielding voluminous amounts of digital data (http://data.gov/). The State Department has taken the unprecedented step of making every single public statement by the Secretary of State Hillary Clinton, both inside and outside the country, available for download on their website (http://www.state.gov/). We compiled a data set of every public statement made by Secretary Clinton since being appointed using a program called Sitesucker (http://sitesucker.us/). Today our *Clinton Dataset* consists of 2,211 speeches, and it grows every week. It is 178 MB large. Even though the source is publicly available, most social scientists attempting to analyze Clinton's statements are only using very small samples [7, 8]. In contrast, we will analyze all of her statements, and we will establish an ongoing system for integrating new material as it becomes available.

3) Economic History: The second opportunity aims to explain the extent to which network effects explain the long-term increase in personal bankruptcy from less than 1 per 100,000 persons under the temporary laws passed in the 19th century, to 3 per 10,000 in 1938, to 6 per 1,000 in 2005. Attempts to quantify the importance of network effects have been limited to modern cross-sectional or short panel studies. We will fill the gap in this literature by using textual analysis to construct independent variables measuring (1) the frequency with which consumer credit and bankruptcy were mentioned in the popular press and (2) whether the expansion of consumer credit and bankruptcy are portrayed in a pro-debtor or pro-creditor manner. The US Bankruptcy Dataset has not yet been collected. We will extract the data from searchable electronic texts of the popular press available from American Periodicals Series Online and searchable digital versions of newspapers including the Atlantic Constitution, the Chicago Tribune, the Los Angeles Times, the Wall Street Journal, the Washington Post, the New York Times, and Harper's Weekly.

Computing Environment

Development of the ASSANA Methodology for a personal computing environment will use commercially available and open source technology and will facilitate very broad adoption. But for very large data sets, the PC environment is still quite slow. Many new users of text-mining software quickly become frustrated at how long it takes to run analyses, especially as their datasets get larger and more complex. In a high performance computing (HPC) environment processing is much faster. HPC environments have increased speed and capacity: a typical computationally intensive analysis using a PC that could take up to one week to complete, might take as little as three hours in an HPC environment. An HPC environment is also capable of taking input in real time and continuously processing it to provide near real-time analysis of data with a lag of as little as one day.

Refinement

After completion of the Clinton Dataset and US Bankruptcy Dataset studies, we will step back to reflect on what we have learned from each of the four approaches; what research questions were we able to answer (or not answer), and what approaches worked best for inductive or deductive questions. We will compare the amount of data each approach was able to analyze and the amount of time each approach took, in order to create a useable method and tool social scientists can then implement when using their own datasets. We will integrate additional literature into our thinking, and reflect on feedback from conference presentations and publications about the methodology. We will incorporate that reflection into a refined methodology and tool, which will prepare us for the final phase of the project, where we will focus on dissemination of the methodology to the larger scholarly community.

IMPACT

The ASSANA project will have a substantial impact on the ability of US researchers to make better use of the available digital data being produced through US government funding. It will accelerate scientific advancement through the engagement in collaborative large-N and comparative studies. Through panel presentations and workshops we will disseminate and train social scientists on the various research software available and the best uses for them. In addition, there is already an increasing realization within the national security community that sentiment analysis is a critical tool in protecting US borders and population.

Because the ASSANA methodology will contribute to how sentiment analysis is used to answer increasingly subtle questions about large-scale text-based data, it is expected to contribute to enhancing national security and assist the State Department in its Public Diplomacy efforts.

Even beyond the impact on scholarship and national security, the Washington Post is using these techniques to discern patterns in text-based data, such as the coverage about espionage and more recently, an article (http://www.washingtonpost.com/wp-srv/special/nation/guns/) about the Hidden Life of Guns, Sunday Nov 21.

REFERENCES

1. F. Berman and H.E. Brady, *Workshop on Cyberinfrastructure for the Social and Behavioral Sciences: Final Report*, 2005.

2. M. Gutmann, “Rebuilding the Mosaic: Fostering Research in the Social, Behavioral, and Economic Sciences at the National Science Foundation in the Next Decade,” *National Science Foundation, Directorate for Social, Behavioral and Economic Sciences*, October 2011, NSF 11-086

3. S.R. Corman, A. Trethewey, and B. Goodall, *A 21st Century Model for Communication in the Global War of Ideas: From Simplistic Influence to Pragmatic Complexity*, 2007.

4. K.M. Lord and M. Lynch, *America’s Extended Hand: Assessing the Obama Administration's Global Engagement Strategy*, 2010.

5. J. Nye, *Soft Power: The Means to Success in World Politics*, Public Affairs, 2005.

6. H. Lasswell, “The theory of political propaganda,” *The American Political Science Review*, vol. 21, 1927, pp. 627-631.

7. S. Major, “Sharpening Our Plowshares: Applying the Lessons of Counterinsurgency to Development and Humanitarian Aid,” *SSRN eLibrary*, 2010.

8. R. Winthrop, “Punching Below its Weight: The U.S. Government Approach to Education in the Developing World,” *SSRN eLibrary*, 2010.