Document mining in data journalism

A research proposal

1. Introduction

In the fall of 2010, the WikiLeaks organization released nearly 400,000 text documents relating to the conduct of US armed forces and independent security contractors during the war in Iraq[1]. Since that time, specialized investigative "data journalists" have reported on what was contained in this vast deposit of documents, which included startling information regarding civilian casualties, friendly fire incidents, and observed breaches of protocol. My proposed research, in brief, asks how data journalists "mine" such deposits, how they seek information to support or refute prior findings, or how they come to discover unanticipated yet newsworthy stories hiding in the data.

The motivation for this research proposal reflects the growing trend of large collections of emails, reports, and other documents being "dumped", "leaked", released, or declassified by corporations, government agencies, and other organizations, such as WikiLeaks. Fellow journalists and the news-reading public deserve transparency when it comes to the methods of data journalists who investigate these collections[2]. Additionally, developers of data analysis applications require a better understanding of journalists as potential users.

Aside from being a consumer of news media, I have no background in journalism. Rather I represent the interests of application developers, and I am sensitive to

[1] See Wikipedia on WikiLeaks and The New York Times on WikiLeaks

[2] ReportersLab interviews C.W. Anderson, assistant professor of communications at CUNY

theoretical concepts relating to the design of tools for supporting data analysis. While this theoretical repertoire is useful for understanding low-level perceptual activity contributing to how individuals interact with data displays [2, 12], or for understanding high-level domain-agnostic abstractions relating to *information foraging* and *sense-making* [1, 9], I am faced with a gap in the middle. How can I characterize the data analysis process within the context of a domain such as journalism? Moreover, how does this process play out with a specific type of data, in this case being collections of text documents? Thus I seek a *middle-level* theory to explain this process.

2. Research questions

My predominant research question asks: What is document mining? That is, how do journalists conduct data analysis when faced with more documents than they could possibly read in a year, let alone in time to meet a deadline?

Several additional questions follow from this: namely, what constraints do these journalists face in the process of their work? What tools do they use and how do they use them? Do they collaborate with other people, and if so, how do they collaborate?

Finally, this work will determine how document mining compares to other analytical processes in data journalism, when the data is comprised of numbers rather than text documents, which could include large financial databases or historical measurements. I will also make comparisons between document mining and other processes of investigative journalism, as well as with processes of data analysis characterized in other domains, such as business intelligence and law enforcement.

3. Research context

Journalists engaging in document mining are found in newsrooms around the world. Unfortunately, like many busy professionals, they are often working under tight deadlines, and have little time to participate in academic research. However, I am lucky enough to know someone “on the inside”.

Jonathan Stray is a computer-scientist-turned-journalist now employed by the Associated Press (AP), based out of New York City. Working in collaboration with my research group, he has developed *Overview*, a robust data visualization application for document mining, recently made available as a free download on the AP’s website[3]. He is currently pitching *Overview* to journalists via conferences, workshops, and social media. Buzz surrounding *Overview* is starting to grow in the data journalism community.

Mr. Stray is our gatekeeper to research participants, as his potentially useful application provides an incentive for journalists participate in our research. As a result, we have an opportunity to satisfy two research goals: (1) assess whether *Overview* is usable and useful, as well as how it fits within existing document mining workflows; and (2) characterize the process and context of document mining, with and without this new application. While this proposal focuses on the latter goal, data collection corresponding to both goals will occur simultaneously. Furthermore, it is my intent that in working toward the second goal, my findings will contribute to the future development of *Overview* and other applications like it.

[3] See overview.ap.org/

Due to the distributed nature of this research, logistical constraints will keep me from visiting individual journalists in newsrooms. Thus my data collection will occur at a distance, over the phone and online.

4. Methodology

A need to characterize the process of document mining among journalists necessitates a grounded theory approach [3]. This approach is in turn informed by an interpretivist, symbolic interactionist theoretical perspective and a constructionist epistemology [4]. That is, I intend to focus on the language used by journalists to describe this process, and construct a shared interpretation of this process based on interactions with research participants and the data they generate.

The constant comparative method of grounded theory will allow me to flexibly make comparisons between the process of document mining with other journalistic processes, as well as processes relating to data analysis in other domains. Comparisons will also be made between journalists, between newsrooms, and over periods of time. For instance, I will be comparing the process of document mining both before and after the introduction of *Overview*, the new visualization tool.

A further justification for the use of a grounded theory methodology is that my initial research questions are not theoretically deduced hypotheses. Rather, my questions are informed by sensitizing concepts and assumptions held within my domain. It is these sensitizing concepts that allowed me to frame the data collection methods, particularly a preliminary set of interview questions.

These sensitizing concepts include the notion that data analysis, document mining being an instance of which, occurs in stages. Data analysis may involve stages of hypothesis generation, each necessitating an exploration of the data without a particular set of questions in mind, save perhaps “What’s going on here?”. Other times there may be stages of hypothesis validation, where the goal is to support or refute prior evidence. These stages necessitate a directed search within a subset of the data, or a comparison between individual items or documents. Individuals may or may not engage in both types of stages during the course of a single investigation.

The products of data analysis are also among my sensitizing concepts. These products include “Eureka!” moments of insight, serendipitous discoveries, and both optimal and suboptimal solutions to closed- and open-ended problems. Admittedly, these products of analysis are ill-defined constructs, and it will be necessary to attain our research participants’ interpretations of their meaning, as well as their native terminology.

Finally, these sensitizing concepts include the disentanglement of an individual’s expertise. By this I mean that an analyst may have expertise within a domain, expertise using specific analytical tools or techniques, and/or expertise regarding the data, its semantics and its provenance.

In my field of research, there exist several precedents for the use of grounded theory, or at least the use of methods inspired by a grounded theory methodology. The methodology has informed prior work which has characterized the data analysis processes of professionals in other domains, including architecture [14] and national security and intelligence [7]. There also exists a “grounded evaluation” technique for determining the effectiveness of visualization software when deployed in target contexts of use [6]. Both uses of grounded theory methods serve as inspiration for my proposed research.

4.1. Data collection methods and sources

Primary: My primary data collection method will be intensive, open-ended interviews with journalists. These interviews will be teleconferences or group Skype chats. Both Mr. Stray (at the AP in New York), and myself (at UBC) will have questions to ask interviewees, with his questions pertaining to the usability and utility of *Overview*. Audio from these interviews will be recorded for later transcription.

Following the methodology of grounded theory, I will not specify the number of interviews that I plan to conduct a priori. The final number of interviews will depend on how much theoretical sampling is required before achieving data saturation, the point where no new categories emerge; I will return to this point in the following section. The number of interviews will also depend on how many journalists download and use *Overview*, and among those, how many express a willingness to participate in interviews. Ideally, I would like to perform multiple interviews with each journalist in order to make comparisons over time, as their processes vary or change over time, before, during, and after using the new tool. However, this is an unrealistic plan. As mentioned above in Section 3, these journalists will often be conducting their investigation and writing their story under a tight deadline, and will likely only have time to commit to an intensive interview after the story is written. As a result, I will rely on secondary data collection methods, such as follow-up email exchanges, to fill in some of the gaps. These methods are discussed in greater detail below.

Regarding the content of these interviews, I plan to keep the number of initial questions small. I have prepared a list of interview foci with a small set of representative questions for each, composed according to guidelines for open-ended interviews [5] and for interviews conducted in the context of a grounded theory study [3]. These foci correspond with the research questions mentioned above in Section 2. In particular, I will attempt to ground the interview in the interviewee's example of document mining, one drawn from their prior experience. This will invite comparisons with other journalistic processes, as well as comparisons between their processes before and after their initial use of *Overview*.

Many questions are redundant and cross-referential, a deliberate choice, as I have no intention to ask all or even most of them in a single interview. An answer to one open-ended question is expected to answer many of the others; these foci and questions are more so a checklist than a script. I also expect this list of foci and questions to change as I conduct interviews, as a result of theoretical sampling and the possibility that early interviews will illuminate unanticipated themes and concepts.

Secondary: I plan to complement the interviews by eliciting texts and other information from journalists that I interview. Follow-up questions will be asked via email. I will also request copies of the notes journalists take during the course of their investigation. I expect that in many cases, journalists will be taking notes regardless of whether or not I ask to see them. I will also request information regarding the data, such as how many documents are contained in the document collection being investigated, how they tend to vary from one another, how many were read or skimmed during the course of their investigation, how many were discarded or ignored, as well as why individual documents were read, skimmed, ignored, or discarded. In cases where these documents are publicly available, I will examine the documents as well. Screenshots or pictures of annotated documents, journalist notes, and other analytical artifacts, such as spreadsheets and data visualizations, will also be requested.

Realizing the value of found data [13], I will also collect several extant texts. In particular, I will collect the stories journalists write as a result of their investigations. I will examine the extent to which the document mining process is transparent in their articles, allowing for a comparison with their notes and the remarks they make during interviews. Finally, in cases where these stories are published online, I will also collect the reader perspective, via comments and discussion boards.

4.2. Data analysis methods

Data collection and analysis for this project will occur concurrently. Interviews and artifacts collected from journalists will be subject to multiple iterations of coding, each calling upon the constant comparative method, the basis of grounded theory [3]. First, *open coding* will label the data, at the line or paragraph level, using words or short phrases used in the data. Next, tentative categories of codes will be generated, each with an explanatory rationale based on comparisons between code instances, recorded in memos. This will inform subsequent data collection and focused coding. The process of *axial coding* follows, a recoding of the data using the categories constructed. At this point, the process of theoretical sampling will direct me to specific data collection, using different interview foci or artifact collection criteria. As categories become refined and theoretical concepts emerge through the process of *selective coding* and memoing, I will begin to seek theoretical saturation, the point where no unexamined concepts are apparent. At this time, I will begin to construct a mid-level theory of document mining based on the relationships between concepts. This stage will also involve comparisons between my theory and other theories of data analysis, as it occurs in other domains and as it is described at higher levels of abstraction [1, 9].

Triangulation between researchers is highly effective during interpretive analysis [8]. I will share my findings with Mr. Stray. While he and I have different foci and research goals, we will both be engaged in participant interviews, and thus can compare notes. Additionally, his own journalistic expertise can also be called upon throughout the stages of my analysis.

I will also triangulate in terms of methods, in that I will take an alternative approach to thematic coding [11]. This will involve an examination of word frequency, word co-occurrence, key words as used in context, linguistic connectors, and metaphors used. Extant texts and artifacts collected will also be analyzed in terms of their descriptive properties, as well as their intellectual and cultural values [10]. I will then compare the codes produced by these alternative techniques to the codes and categories generated via the grounded theory methods.

5. Outcomes and follow-on work

I anticipate two audiences to which I will report my findings. The first are readers of peer-reviewed academic publications and/or conference proceedings in the fields of human-computer interaction, information visualization, and visual analytics. The second audience for my findings are journalists and the news-reading public, so I plan to report my findings online, either via my own website or in collaboration with the AP.

I hope to apply my findings in the future development of *Overview* and other applications like it. Finally, I anticipate that an examination of what makes *Overview* ultimately successful or unsuccessful will call for a critical inquiry of existing values and standards in journalism, as well as existing theories of data analysis.

References

1. Amar, R. and Stasko, J. T. (2004). A knowledge task-based framework for design and evaluation of information visualizations. In *Proc. IEEE Symp. Information Visualization (InfoVis)*, 143-150.
2. Amar, R., Eagan, J., and Stasko, J. T. (2005). Low-level components of analytic activity in information visualization. In *Proc. IEEE Symp. Information Visualization (InfoVis)*, 111-117.
3. Charmaz, K. (2006). *Constructing Grounded Theory: A Practical Guide through Qualitative Analysis*. Sage Publications Ltd.
4. Crotty, M. (1998). *The Foundations of Social Research*. Sage Publications, Inc..
5. Fontana, A. and Frey, J. (1994). Interviewing: The Art of Science. In *The Handbook of Qualitative Research*, 361-376. Sage Publications.
6. Isenberg, P., Zuk, Collins, T. C., and Carpendale S. (2008). Grounded evaluation of information visualizations. In *Proc. Conf. BEyond time and errors: novel evaLuation methods for Information Visualization (BELIV)*.
7. Kang, Y. A. and Stasko, J. T. (2011). Characterizing the intelligence analysis process: Informing visual analytics design through a longitudinal field study. In *Proc. IEEE Symp. Visual Analytics Science and Technology (VAST)*, 19-28.
8. Mathison, S. (1988). Why Triangulate?. *Educational Researcher 17*, 13.
9. Pirolli, P. (2009). *Information Foraging Theory: Adaptive Interaction with Information*. Oxford University Press, Inc.
10. Prown, J. D. (1982). Mind in Matter: An Introduction to Material Culture Theory and Method. *Winterthur Portfolio* 17, 1-19.
11. Ryan, G. W. and Bernard, H. R. (2003). Techniques to Identify Themes. *Field Methods 15*, 85-109.
12. Shneiderman, B. (1996). The eyes have it: a task by data type taxonomy for information visualizations. In *Proc. IEEE Symp. on Visual Languages*, 336-343.
13. Silverman, D. (2007). *A Very Short, Fairly Interesting and Reasonably Cheap Book about Qualitative Research*. Sage Publications.
14. Tory, M. and Staub-French, S. (2008). Qualitative analysis of visualization: A building design field study. In *Proc. Conf. BEyond time and errors: novel evaLuation methods for Information Visualization (BELIV)*.