Managing and Visualizing Unstructured Big Data

Ananda Mitra
Wake Forest University, USA

INTRODUCTION

This essay expands on the notion of "Big Data" to open up alternative analytic opportunities, on certain components of the data, through a theoretical lens that is mobilized to offer an interpretation and visualization of the information contained in large amounts of data. It is useful first to examine the term "Big Data" in some detail. The term refers to a phenomenon which results from the fact that institutions and individuals are digitizing many different kinds of information leading to an exponential growth of the amount of data can that is being stored in the digital space. First, this expansion relates to the *increase in data points* as more records are added to the corpus of Big Data. Second, the idea of Big Data needs to be considered in terms of the *details that are being digitized*. The notion of Big Data should be considered both in terms of the breadth of the data in terms of number of data points (*amount*) and the depth of the data related to the various fields of information available for each record (*details*). Therefore, Big Data has become an object of analysis for a variety of groups, from academics to marketers, all of whom are interested in understanding how Big Data could provide highly granular voluminous information about people (see, e.g., Mitra, 2014c). Next, it is useful to examine the different categories of information that makes up "Big Data."

Much of Big Data is numeric that is amenable to mathematical analysis. For instance, it is possible to easily count the number of tweets produced by an individual. Such counts offer the opportunity for companies such as Tweeter to offer information about what topics are popular at any moment in time. The segment of Big Data that offers the ease of analysis and visualization has been called "structured" Big Data. There is, however, another vast component of Big Data that does not allow for easy numeric analysis. This segment is made up of the utterances of the people who are self-generating the Big Data by voicing themselves in the digital space. An example of this segment of Big Data is the actual specific tweet produced by an individual or the specific photograph uploaded on photograph sharing spaces. In the case of the tweet, the language of the tweet contains information about attitudes and opinions, just as a photograph offers information about the individual who has captured the picture. This form of data requires a more nuanced and "qualitative" analytic process that would discover the intent of the authors and the meanings of the messages encapsulated in a microblog or picture. This segment of Big Data has been named "unstructured" Big Data. Not only is it difficult to analyze the unstructured Big Data but it is also difficult to visualize the findings of analysis. The qualitative process does not typically produce convenient charts and graphs. The analysis needs to be offered for easier understanding and unstructured Big Data makes this a challenge as well.

This paper offers a theoretical and analytic process to consider ways of analyzing Big Data and visualizing the analysis. To do this, it is important to offer a theoretical basis to consider the elements of unstructured Big Data.

DOI: 10.4018/978-1-5225-2255-3.ch035

BACKGROUND

Perspective on Big Data

The unstructured Big Data can be categorized into three main types. The first set are characterized by short word length where the information is authored by individuals and institutions. This form of the data has sometimes been called "micro-blogs," as reference to blogs that are strictly restricted by the number of words that can be used in the discourse. The most popular example of micro-blogs are the statements produced by the users of the computer program called Tweeter. The second set of unstructured Big Data is an extension of micro-blogs where the restriction on size disappears but all the other characteristics remain intact. This is a situation where a user can generate discourse of significant length and place it in a digital repository. One popular example of this category are "posts" that users upload within their Facebook "profiles." The third category of unstructured Big Data that is worthy of consideration is discourse that users generate in response to specific queries. This form of data is rarely circulated over the Internet, but remains as in-depth lengthy treatise on very specific issues that the user is prompted to elaborate. Much like the second category, this segment of unstructured Big Data is not usually restricted in length. However, there are greater restrictions on the scope of content of this form of data since a majority of this data is generated in response to prompts and questions. A good example of this form of data are responses to open-ended questions used in questionnaires in varieties of data collection projects ranging from measuring political opinions to public health assessments. This three-pronged categorization encompasses the majority of unstructured Big Data with one common characteristic – the data is user-generated. It is therefore useful to consider how to characterize the author of the discourse. I offer two broad categories, which can be considered to be mutually exclusive for most considerations. The two user groups are "institutions" and "individuals" with the differentiating factor being the level of agency of the user.

Institutions often bring the full force of their creative, financial and cultural capital to the creation of digital data, often in the form of "home pages" that populate the digital space. This form of data represents an "authoritative" voice of the powerful and dominant within the public sphere (Foucault, 1991). Such voices carry the ideological baggage of the institutions and their relative position along a continuum of power from the dominant to the oppositional. However, with the availability of the digital tools and the relative ease with which digital data can be produced and circulate it is increasingly possible for some individuals to create such home pages which represented the voice of the relatively powerless individual who would not have the institutional support to present themselves in the public sphere (see, e.g., Mitra and Watts, 2002; Mitra, 2011). Starting with home pages, individuals were able to gain a sense of agency and authorship in the digital space where the individual user-generated data was beginning to become available in the digital space. This tendency expanded rapidly with the development of better tools for digital communication and the ease with which individuals could constantly generate data and place it within networks of other users who were also generating data. Consequently, a new state of empowerment was being achieved by the individuals where their voice was becoming alongside the authoritative voices of the institutions. The corpus of Big Data is thus made up of the institutionally produced and individually authored information. In the remainder of the essay, the focus is on the three categories of Big Data – micro-blogs, posts, and response to open ended questions – produced by individuals. In the next section a specific theoretical approach is suggested to help analyze this form of data.