Chapter 7
Managing and Visualizing Unstructured Big Data

Ananda Mitra
Wake Forest University, USA

ABSTRACT

One of the most common terms that is used in a significant amount of popular and scholarly discussion is big data. As pointed out earlier, the term has a dubious history and different people claim ownership. What remains true of the notion of big data is that it exists. With the increasing rate at which institutions and individuals are digitizing many different kinds of information the amount of data can only go up in volume. 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 information about people.

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."

DOI: 10.4018/978-1-5225-7598-6.ch007

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.

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