Survey of the Different Type of Data Analytics Algorithms

Michael Kevin Hernandez, Boeing, Charleston, SC, USA

ABSTRACT

From as early as 1854 to today, society has been gathering, processing, transforming, modeling and visualizing data to help drive data-driven decisions. The qualitative definition of big data can be defined more conclusively as data that has high volume, velocity, and variety. Whereas, the quantitative definition of big data does vary with respect to time due to the dependence of the time’s technology and processing capabilities. However, making use of that big data to facilitate data-driven decisions, one should employ either descriptive, predictive, or prescriptive analytics. This article has discussed and summarized the advantages and disadvantages of the algorithms that fell under descriptive and predictive analytics. Given the sheer number of the different types of algorithms and the amount of versatile data mining software available sometimes, the best big data analytics results can come from mixing two to three of the mentioned algorithms.

KEYWORDS

Big Data, Data Analytics Algorithms, Descriptive Analytics, Predictive Analytics

INTRODUCTION

Data analytics has existed before 1854, and since then technology has advanced so much that people and companies are collecting more data, from different source and structures, and at high speeds. However, a misconception that has formed is that gathering as much data as possible to increase a company’s competitive advantage is hype, because there will come the point where the gathered information becomes burdensome for a group of people to go through without the aid of some big data analytics and data mining algorithms (Vuori, 2006). Fayyad, Piatetsky-Shapiro, and Smyth (1996) stated that data mining hype could be drawn up due to the positive media attention the topic has garnered. Fox and Do (2013), warned managers to filter out the hype, because if not it can severely influence potential investment decisions. To help filter the hype, Gartner produces the Hype Cycle for different technology spaces, usually displaying the technology in one of five phase: Innovation Trigger (potential acknowledged), Peak of Inflated Expectations (media shares success stories), Trough of Disillusionment (interests fades), Slope of Enlightenment (introducing nth generation of products), and Plateau of Productivity (mainstream adoption) (Schlegel, 2015). As of August 2015, big data is in the “Trough of Disillusionment” phase with about 5-10 years of reaching mainstream adoption. Big data analytics can be broken up into three categories: Descriptive analytics has surpassed mainstream adoption; Predictive analytics is at the top of the “Peak of Inflated Expectations” phase; and Prescriptive Analytics is just entering the “Peak of Inflated Expectations” phase (Schlegel, 2015).

The objective of this paper is to discuss the history of data analytics, help define what makes big data different than conventional data, and types of big data analytics. Finally, this paper will summarize the most common and different big data mining algorithms for both descriptive and

DOI: 10.4018/IJSITA.2017010104

Copyright © 2017, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
predictive analytics. Prescriptive analytics will take ten years to become mainstream (Schlegel, 2015). Thus, this type of analytics will not become the focus of this paper. Also, even though prescriptive analytics deal with optimization, this is too new to a field with little tangential applications in other fields. However, it should be noted that prescriptive analytics has components of both predictive and descriptive analytics.

**HISTORY OF DATA ANALYTICS**

Data analytics has existed before 1854. Snow (1854) had a theory on how cholera outbreaks occur, and he was able to use that theory to remove the pump handle off of a water pump, where that water pump had been contaminated in the summer of 1854. He had set out to prove that his hypothesis on how cholera epidemics originated from was correct, so he then drew his famous spot maps for the Board of Guardians of St. James’ parish in December 1854. These maps were showed in his eventual 2nd edition of his book “On the Mode of Communication of Cholera” (Brody, Rip, Vinten-Johansen, Paneth, & Rachman, 2000; Snow, 1855). As Brody et al. (2000) stated, this case was one of the first famous examples of the theory being proven by data, but the earlier usage of spot maps has existed.

However, the use of just geospatial data analytics can be quite limiting in finding a conclusive result if there is no underlying theory as to why the data is being recorded (Brody et al., 2000). Through the addition of subject matter knowledge and subject matter relationships before data analytics, context can be added to the data for which it can help yield better results (Garcia, Ferraz, & Vivacqua, 2009). In the case of Snow’s analysis, it could have been argued by anyone that the atmosphere in that region of London was causing the outbreak. However, Snow’s original hypothesis was about the transmission of cholera through water distribution systems, the data then helped support his hypothesis (Brody et al., 2000; Snow 1854). Thus, the suboptimal results generated from the outdated Edisonian-esque, which is a test-and-fail methodology, can prove to be very costly regarding Research and Development, compared to the results and insights gained from text mining and manipulation techniques (Chonde & Kumara, 2014).

Snow’s (1854) study was the first well-known case of data analytics. However, the first two cases of text mining originated from Luhn (1953; 1957). Luhn (1953) proposed that there should be a systematic way to assigning key terms to the body of text because if two investigators were to do so manually, they could come up with different key term results. On Luhn (1957), a systematic and automate statistical approach to the 1953 paper was created, which assigned key terms to 1200 documents. Luhn could not conclude which parts of speech (verbs or nouns) made better keywords (1957). However, the English language had synonyms and polynyms, which added complexity to text mining and building key terms to a body of text. Thus, Dumuis, Furnasa, Landauer, Deerwester, and Harshman (1988), built on Luhn (1957) work by proposing the use of latent semantic analysis for text mining. Term-weighted approaches to automated text key term indexing were introduced by Salton and Buckley (1988). Since then the field of text mining has grown.

From analyzing deaths in 1854 to solving the issue of document indexing via assigning key terms, previous research has helped pave the way for what is considered as data and text analytics. Previously the analysis of that many data points was considered big data.
Enhancing Decision Support Systems with Spatial Capabilities
[www.igi-global.com/chapter/enhancing-decision-support-systems-spatial/36832?camid=4v1a](www.igi-global.com/chapter/enhancing-decision-support-systems-spatial/36832?camid=4v1a)

A New Metric of Validation for Automatic Text Summarization by Extraction
[www.igi-global.com/article/a-new-metric-of-validation-for-automatic-text-summarization-by-extraction/209458?camid=4v1a](www.igi-global.com/article/a-new-metric-of-validation-for-automatic-text-summarization-by-extraction/209458?camid=4v1a)