Summary of "Beyond the hype: Big data concepts, methods and analytics"

D'Ambrosi Denis

October 9, 2023

This week's article provides an overview to the world of Big Data by elaborating on its definition, discussing the main challenges involved in this field and introducing a little bit of taxonomy of its major practical applications, along with the relative computational techniques.

First of all, the authors clarify a common misconception about Big Data, rooted directly in its name: size is not the only defining feature of this discipline. Classifying data (and its mining strategies) only based on the sheer amount of bytes it occupies on an hard drive is a simplistic attitude that does not reflect all the nuances of this field: information may be registered with highly different degrees of velocity (rate of generation and analysis of data), variety (number of possible formats in which it may be retrieved), veracity (the level of reliability linked to the information received), variability (a metric that depends on the fluctuations of the previously introduced velocity) and value (the quantity of actual useful insights gained from data analysis) are all challenges that have to be tackled within the Big Data world, along with the actual volume of information. Furthermore, such metrics are meant to evolve as we head into an increasingly data-driven economy, so it would be pointless to try to identify Big Data techniques exclusively with regards of the size of their subject.

After discussing more abstractly the challenges of this field, the authors present a possible classification of the various Big Data domains based on the nature of the information processed. They highlight five main areas of application: text analytics, audio analytics, video analytics, social media analytics and predictive analysis.

Text analytics' main objective is extracting information from unstructured textual data. Some relevant techniques to this sub-field are: information extraction (that aims at recognising and classifying single words and relationships between words in a written document), text summarization (which can either exploit statistical or semantical techniques to shorten the content of a text), question answering (that have to understand questions and provide answers in natural language) and sentiment analysis (which aims at classifying a textual information based on the subjective opinions of its writer).

Audio analytics' end is to extract information from audio data (generally recordings of human speech). Generally speaking, the information is mined

either by translating the audio file into a written transcript and then applying text analytics techniques or by processing the extracted phonemes directly.

Video analytics techniques have recently becomed incredibly relevant as we have never had as much heterogenic video data sources as today (social media, CCTV cameras, common visual medias, advertising, etc...). Real world application of such systems range from real-time surveillance to market analysis, but their taxonomy is mainly based on the physical location in which the computation takes place: if the processing architecture features a main computational point it is called "server-based", while if each node capable of recording video is also responsible of mining through it, than the system is called "edge-based".

Social Media are a never-ending source of information produced by a very heterogenic group of users: processing all this data may be cumbersone, but it allows scientists to extract analytics based both on single pieces of contents and on social structures and groups (such techniques are at the core of modern recommendation systems).

Finally, predictive analysis aims at producing predictions on future outcomes based on historical data. This requires to infere patterns from input data in order to guess probable features of unseen datapoints. Such techniques have to face many challenges, among which the authors have pointed out heteogeneity, noise accumulation, spurious correlations and incidental endogeneity.