Information Fusion 28 (2016) 45–59

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Social big data: Recent achievements and new challenges

Gema Bello-Orgaz [a], Jason J. Jung [b,*], David Camacho [a]

[a] Computer Science Department, Universidad Autónoma de Madrid, Spain
[b] Department of Computer Engineering, Chung-Ang University, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Available online 28 August 2015

Keywords:
Big data
Data mining
Social media
Social networks
Social-based frameworks and applications

ABSTRACT

Big data has become an important issue for a large number of research areas such as data mining, machine learning, computational intelligence, information fusion, the semantic Web, and social networks. The rise of different big data frameworks such as Apache Hadoop and, more recently, Spark, for massive data processing based on the MapReduce paradigm has allowed for the efficient utilisation of data mining methods and machine learning algorithms in different domains. A number of libraries such as Mahout and SparkMLib have been designed to develop new efficient applications based on machine learning algorithms. The combination of big data technologies and traditional machine learning algorithms has generated new and interesting challenges in other areas as social media and social networks. These new challenges are focused mainly on problems such as data processing, data storage, data representation, and how data can be used for pattern mining, analysing user behaviours, and visualizing and tracking data, among others. In this paper, we present a revision of the new methodologies that is designed to allow for efficient data mining and information fusion from social media and of the new applications and frameworks that are currently appearing under the "umbrella" of the social networks, social media and big data paradigms.

1. Introduction

Data volume and the multitude of sources have experienced exponential growth, creating new technical and application challenges; data generation has been estimated at 2.5 Exabytes (1 Exabyte = 1.000.000 Terabytes) of data per day [1]. These data come from everywhere: sensors used to gather climate, traffic and flight information, posts to social media sites (Twitter and Facebook are popular examples), digital pictures and videos (YouTube users upload 72 hours of new video content per minute [2]), transaction records, and cell phone GPS signals, to name a few. The classic methods, algorithms, frameworks, and tools for data management have become both inadequate for processing this amount of data and unable to offer effective solutions for managing the data growth. The problem of managing and extracting useful knowledge from these data sources is currently one of the most popular topics in computing research [3,4].

In this context, **big data** is a popular phenomenon that aims to provide an alternative to traditional solutions based on databases and data analysis. Big data is not just about storage or access to data; its solutions aim to analyse data in order to make sense of them and exploit their value. Big data refers to datasets that are terabytes to petabytes (and even exabytes) in size, and the massive sizes of these datasets extend beyond the ability of average database software tools to capture, store, manage, and analyse them effectively.

The concept of big data has been defined through the **3V model**, which was defined in 2001 by Laney [5] as: *"high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making"*. More recently, in 2012, Gartner [6] updated the definition as follows: *"Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization"*. Both definitions refer to the three basic features of big data: **V**olume, **V**ariety, and **V**elocity. Other organisations, and big data practitioners (e.g., researchers, engineers, and so on), have extended this 3V model to a 4V model by including a new "V": **V**alue [7]. This model can be even extended to 5Vs if the concepts of **V**eracity is incorporated into the big data definition.

Summarising, this set of ∗V-models provides a straightforward and widely accepted definition related to what is (and what is not) a big-data-based problem, application, software, or framework. These concepts can be briefly described as follows [5,7]:

- ***V**olume*: refers to large amounts of any kind of data from any different sources, including mobile digital data creation devices and digital devices. The benefit from gathering, processing, and analysing these large amounts of data generates a number

* Corresponding author. Tel.: +821020235863.
E-mail addresses: gema.bello@uam.es (G. Bello-Orgaz), j2jung@gmail.com, j2jung@yahoo.com, j3ung@cau.ac.kr (J.J. Jung), david.camacho@uam.es (D. Camacho).

http://dx.doi.org/10.1016/j.inffus.2015.08.005
1566-2535/

of challenges in obtaining valuable knowledge for people and companies (see Value feature).

- ***V**elocity*: refers to the speed of data transfers. The data's contents are constantly changing through the absorption of complementary data collections, the introduction of previous data or legacy collections, and the different forms of streamed data from multiple sources. From this point of view, new algorithms and methods are needed to adequately process and analyse the online and streaming data.
- ***V**ariety*: refers to different types of data collected via sensors, smartphones or social networks, such as videos, images, text, audio, data logs, and so on. Moreover, these data can be structured (such as data from relational databases) or unstructured in format.
- ***V**alue*: refers to the process of extracting valuable information from large sets of social data, and it is usually referred to as *big data analytics*. Value is the most important characteristic of any big-data-based application, because it allows to generate useful business information.
- ***V**eracity*: refers to the correctness and accuracy of information. Behind any information management practice lie the core doctrines of data quality, data governance, and metadata management, along with considerations of privacy and legal concerns.

Some examples of potential big data sources are the Open Science Data Cloud [8], the European Union Open Data Portal, open data from the U.S. government, healthcare data, public datasets on Amazon Web Services, etc. **Social media** [9] has become one of the most representative and relevant data sources for big data. Social media data are generated from a wide number of Internet applications and Web sites, with some of the most popular being Facebook, Twitter, LinkedIn, YouTube, Instagram, Google, Tumblr, Flickr, and WordPress. The fast growth of these Web sites allow users to be connected and has created a new generation of people (maybe a new kind of society [10]) who are enthusiastic about interacting, sharing, and collaborating using these sites [11]. This information has spread to many different areas such as everyday life [12] (e-commerce, e-business, e-tourism, hobbies, friendship, ...), education [13], health [14], and daily work.

In this paper, we assume that *social big data* comes from joining the efforts of the two previous domains: *social media* and *big data*. Therefore, **social big data** will be based on the analysis of vast amounts of data that could come from multiple distributed sources but with a **strong focus on social media**. Hence, social big data analysis [15,16] is inherently interdisciplinary and spans areas such as data mining, machine learning, statistics, graph mining, information retrieval, linguistics, natural language processing, the semantic Web, ontologies, and big data computing, among others. Their applications can be extended to a wide number of domains such as health and political trending and forecasting, hobbies, e-business, cybercrime, counterterrorism, time-evolving opinion mining, social network analysis, and humanmachine interactions. The concept of *social big data* can be defined as follows:

"Those processes and methods that are designed to provide sensitive and relevant knowledge to any user or company from social media data sources when data sources can be characterised by their different formats and contents, their very large size, and the online or streamed generation of information."

The gathering, fusion, processing and analysing of the big social media data from unstructured (or semi-structured) sources to extract value knowledge is an extremely difficult task which has not been completely solved. The classic methods, algorithms, frameworks and tools for data management have became inadequate for processing the vast amount of data. This issue has generated a large number of open problems and challenges on social big data domain related to different aspects as knowledge representation, data management, data processing, data analysis, and data visualisation [17]. Some of these challenges include accessing to very large quantities of unstructured data (management issues), determination of how much data is enough for having a large quantity of high quality data (quality versus quantity), processing of data stream dynamically changing, or ensuring the enough privacy (ownership and security), among others. However, given the very large heterogeneous dataset from social media, one of the major challenges is to identify the valuable data and how analyse them to discover useful knowledge improving decision making of individual users and companies [18].

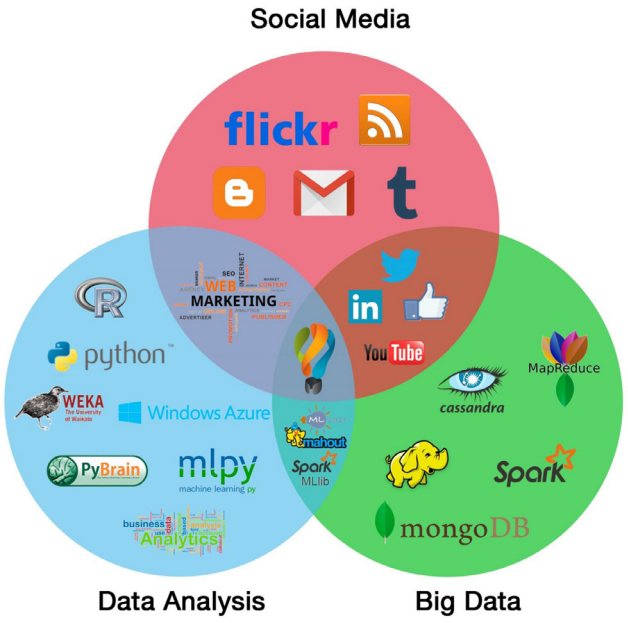

Fig. 1. The conceptual map of Social BigData.

In order to analyse the social media data properly, the traditional analytic techniques and methods (**data analysis**) require adapting and integrating them to the new big data paradigms emerged for massive data processing. Different big data frameworks such as Apache Hadoop [19] and Spark [20] have been arising to allow the efficient application of data mining methods and machine learning algorithms in different domains. Based on these big data frameworks, several libraries such as Mahout [21] and SparkMLib [22] have been designed to develop new efficient versions of classical algorithms. This paper is focused on review those new methodologies, frameworks, and algorithms that are currently appearing under the big data paradigm, and their applications to a wide number of domains such as e-commerce, marketing, security, and healthcare.

Finally, summarising the concepts mentioned previously, Fig. 1 shows the conceptual representation of the three basic social big data areas: **social media** as a natural source for data analysis; **big data** as a parallel and massive processing paradigm; and **data analysis** as a set of algorithms and methods used to extract and analyse knowledge. The intersections between these clusters reflect the concept of mixing those areas. For example, the intersection between big data and data analysis shows some machine learning frameworks that have been designed on top of big data technologies (Mahoot [21], MLBase [23,24], or SparkMLib [22]). The intersection between data analysis and social media represents the concept of current Web-based applications that intensively use social media information, such as applications related to marketing and e-health that are described in Section 4. The intersection between big data and social media is reflected in some social media applications such as LinkedIn, Facebook, and Youtube that are currently using big data technologies (MongoDB, Cassandra, Hadoop, and so on) to develop their Web systems.