Chapter 33
Big Data Analytics and Mining for Knowledge Discovery

Carson K. Leung
https://orcid.org/0000-0002-7541-9127
University of Manitoba, Canada

ABSTRACT

Big data analytics and mining aims to discover implicit, previously unknown, and potentially useful information and knowledge from big data sets that contain huge volumes of valuable veracious data collected or generated at a high velocity from a wide variety of rich data sources. Among different big data analytic and mining tasks, this chapter focuses on frequent pattern mining. By relying on the MapReduce programming model, researchers only need to specify the "map" and "reduce" functions to discover (organizational) knowledge from (i) big data sets of precise data in a breadth-first manner or depth-first manner and/or from (ii) big data sets of uncertain data. Such a big data analytics process can be sped up by focusing the mining according to the user-specified constraints that express the user interests. The resulting (constrained or unconstrained) frequent patterns mined from big data sets provide users with new insights and a sound understanding of users' patterns. Such (organizational) knowledge is useful is many real-life information science and technology applications.

INTRODUCTION

Progresses in information science and technology have enabled the collection and generation of huge volumes of valuable data—such as streams of banking, financial, marketing, organizational, and transactional data—at a high velocity from a wide variety of rich data source in various real-life business, engineering, education, healthcare, hospitality and tourism, scientific, as well as social applications and services in government, organizations and society. These *big data* (Madden, 2012; Leung, 2015; Bellatreche et al., 2019) may be of different levels of veracities (e.g., precise data, imprecise and uncertain data) and/or of a variety of types or formats (e.g., structured data in relational databases; semi-structured data in extensible markup language (XML) or JavaScript object notion (JSON) format stored in document-oriented or graph databases; unstructured data in images, audios and videos). Embedded in the big data

DOI: 10.4018/978-1-6684-3662-2.ch033

is implicit, previously unknown, and potentially useful information and knowledge. However, the big data come with volumes beyond the ability of commonly-used software to capture, manage, and process within a tolerable elapsed time. Hence, new forms of information science and technology—such as *big data analytics and mining for knowledge discovery*—are needed to process and analyze the big data so to as enable the enhanced decision making, insight, and process optimization. For instance, the discovery of organizational knowledge (e.g., common customer complaints, main causes of employee turnover, sets of popular merchandise items in shopping carts)—via techniques like big data analysis, statistics, and business analytics—helps reveal important patterns about an organization. This organizational knowledge helps executive and management teams of the organization to get a better understanding of the organization so that they could make better use of human resources and technology, focus more on education and growth, keep customers top of mind, and further improve quality of services and products. To a further extent, the discovery of organizational knowledge and its subsequent actions help the organization to meet goals, gain competitive advantage, and ultimately ensure sustainability, organizational growth and development.

Over the past two decades, algorithms have been proposed for various big data analytics, mining and knowledge discovery—including clustering (which groups similar data together), classification (which categorizes groups of similar data), outlier detection (which identifies anomalies), and frequent pattern mining (which discovers interesting knowledge in the forms of frequently occurring sets of merchandise items or events). Many of these algorithms use the *MapReduce* model—which mines the search space with distributed or parallel computing (Shim, 2012). Among different big data analytics and mining tasks, this chapter focuses on applying the MapReduce model to big (organizational) data for the discovery of frequent patterns.

BACKGROUND

Since the introduction of the research problem of *frequent pattern mining* (Agrawal, Imieliński, & Swami, 1993), numerous algorithms have been proposed (Hipp, Güntzer, & Nakhaeizadeh, 2000; Ullman, 2000; Ceglar & Roddick, 2006; Aggarwal, Bhuiyan, & Al Hasan, 2014; Leung et al., 2017c). Notable ones include the classical Apriori algorithm (Agrawal & Srikant, 1994) and its variants such as the Partition algorithm (Savasere, Omiecinski, & Navathe, 1995). The Apriori algorithm uses a level-wise breadth-first bottom-up approach with a candidate generate-and-test paradigm to mine frequent patterns from transactional databases of precise data. The Partition algorithm divides the databases into several partitions and applies the Apriori algorithm to each partition to obtain patterns that are locally frequent in the partition. As being locally frequent is a necessary condition for a pattern to be globally frequent, these locally frequent patterns are tested to see if they are globally frequent in the databases. To avoid the candidate generate-and-test paradigm, the tree-based FP-growth algorithm (Han, Pei, & Yin, 2000) was proposed. It uses a depth-first pattern-growth (i.e., divide-and-conquer) approach to mine frequent patterns using a tree structure that captures the contents of the databases. Specifically, the algorithm recursively extracts appropriate tree paths to form projected databases containing relevant transactions and to discover frequent patterns from these projected databases.

In various real-life business, engineering, healthcare, scientific, and social applications and services in modern organizations and society, the available data are not necessarily *precise* but *imprecise or uncertain* (Leung, 2014; Leung, MacKinnon, & Tanbeer, 2014; Cheng et al., 2019; Rahman, Ahmed, &