Big Data Visualization of Association Rules and Frequent Patterns

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

INTRODUCTION

Big Data and *machine learning* are driving Industry 4.0, which is also known as the Fourth Industrial Revolution. Note that the (First) Industrial Revolution transformed manual production to machine production from the late 18th to mid-19th century. The Technological Revolution, which is also known as the Second Industrial Revolution, further industrialized and modernized the industry from the late 19th century to early 20th century through technological advancements and standardization, installations of extensive railroad and telegraph networks, as well as electrification. The Digital Revolution, which is also known as the Third Industrial Revolution, shifted from mechanical and analogue electronic technology to digital electronics, computing and communication technologies in the late 20th century. Now, Big Data have become one of the greatest sources of power in the 21st century, and they have become a critical part of the business world and daily life. In the current era of Big Data, numerous rich data sources are generating huge volumes of a wide variety of valuable data at a high velocity. These Big Data can be of different levels of veracity: They are precise, whereas some others are imprecise and uncertain. Embedded in these Big Data are implicit, previously unknown and potentially useful information and knowledge. This calls for data science, which makes good use of Big Data mining and analytics, machine learning, mathematics, statistics, visualization, and related techniques to manage, mine, analyze and learn from these Big Data to discover and visualize hidden gems. This, in turn, may maximize the citizens' wealth and/or promote all society's health. As one of the important Big Data mining and analytics tasks, frequent pattern mining aims to discover interesting knowledge in the forms of frequently occurring sets of merchandise items or events. For example, patterns discovered from business transactions may help reveal shopper trends, which in turn enhances inventory, minimizes customers' cost, and maximizes citizens' wealth. As another example, patterns discovered from health records may help reveal important relationships associated with certain diseases, which in turn leads to improve and promote all society's health. To enable users to get better understanding of the discovered patterns in a comprehensive manner, several data visualization and visual analytics tools have been proposed. This encyclopedia article covers *Big Data visualization* with focus on *visualization of association rules and frequent patterns*.

BACKGROUND

Since the introduction of the research problem of *frequent pattern mining* (Agrawal et al., 1993), numerous algorithms have been proposed (Hipp et al., 2000; Ullman, 2000; Ceglar & Roddick, 2006; Aggarwal et al., 2014; Leung et al., 2017b; Alam et al., 2021; Chowdhury et al., 2022). Notable ones include the

DOI: 10.4018/978-1-7998-9220-5.ch075

classical Apriori algorithm (Agrawal & Srikant, 1994) and its variants such as the Partition algorithm (Savasere et al., 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 et al., 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.

For different 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 et al., 2014; Rahman et al., 2019; Davashi, 2021; Roy et al., 2022). Examples include sensor data and privacy-preserving data (Leung et al., 2019a; Olawoyin et al., 2021; Pang & Wang, 2021; Jangra & Toshniwal, 2022). Over the past decade, several algorithms have been proposed to mine and analyze these uncertain data. The tree-based UF-growth algorithm (Leung et al., 2008c) is an example.

Moreover, it is not unusual for users to have some phenomenon in mind. For example, a manager in an organization is interested in some promotional items. Hence, it would be more desirable if data mining algorithms return only those patterns containing the promotional items rather than returning all frequent patterns, out of which many may be uninteresting to the manager. It leads to *constrained mining*, in which users can express their interests by specifying constraints and the mining algorithm can reduce the computational effort by focusing on mining those patterns that are interesting to the users.

As we move into the new era of Big Data, these Big Data can be characterized by at least 7Vs—namely, volume, value, velocity, variety, veracity, validity, and visibility. Specifically, volume refers to the huge quantity of Big Data. Value refers to the usefulness of Big Data, as well as information and knowledge discovered from the data. Velocity refers to the rapid rate at which the Big Data are generated or collected. Variety refers to differences in data types (e.g., structured, semi-structured, and/or unstructured data), data contents, data formats (e.g., alphanumeric data, relational data, transactional data, XML, JSON, audio, video), and/or data sources. Veracity refers the quality or trustworthiness of Big Data, especially whether the data are precise vs. imprecise (or uncertain). Validity refers the interpretation of Big Data and of information/knowledge discovered from the data. Visibility refers the visualization of data and of information/knowledge discovered from the data.

Big Data mining and analytics (Leung, 2022) mainly focuses on the first 5V's listed above as it aims to extract valuable information and knowledge from huge volumes of a wide variety of veracious but valuable data that are generated and collected at a high velocity. In contrast, *Big Data visualization* mainly focuses on the last 2V's listed above as it aims to visualize and interpret the valuable data and the information/knowledge that can be discovered from the data. Hence, once Big Data are mined and analyzed to discover valuable information and knowledge (e.g., association rules and frequent patterns), Big Data visualization visualizes these association rules and frequent patterns in a fashion that is comprehendible to users.

Big Data mining and analytics can be executed to discover interesting information and knowledge (e.g., association rules, frequent patterns) from the Big Data. Usually, the discovered knowledge is