Chapter 5
Data Linkage Methods for Big Data Management in Industry 4.0

Onur Doğan
Istanbul Technical University, Turkey

ABSTRACT

In recent years, the use of various digital devices that continuously generate massive amounts of heterogeneous, structured or unstructured data has increased. In parallel to generation, data collection, storage, and analysis technologies have developed. Big data sources have a variety of data quality. Preparing and clearing data is one of the first step of mining big data. It is often important to address the full data set found in different data sources to achieve the right result. Various techniques have been used to increase the accuracy of the data comparison. Deterministic and probabilistic linkage algorithms are the two main techniques used in literature. They have different steps to reach qualified and integrated results. To easily interpret the results of the linkage algorithm, a confusion matrix can be used. Measurements such as sensitivity, specificity, positive predictive value, negative predictive value, false positive rate, and false negative rate, are considered to evaluate output quality.

INTRODUCTION

Big data refers to information and communication technologies that process large amounts of data in order to reach appropriate information to make quick decisions. The volume of data may change from a few terabytes to many petabytes of data. The digital universe expects to grow the volume of data to around 44 zettabytes

DOI: 10.4018/978-1-5225-5137-9.ch005

(40 trillion gigabytes) by 2020. Thanks to embedded systems in internet of things, it is expected that in 2020 volume of data will grow by 10% (Sharma). The use of various digital devices that continuously generate massive amounts of heterogeneous, structured or unstructured data.

Due to huge amounts of data, big data phenome forces many changes in businesses and other organizations, many struggle just to manage the massive data sets and non-traditional data structures that are typical of big data. Managing big data brings together old and new technologies and practices (Russom, 2013). Big data management as a hybrid of old and new best practices, skills, teams and data types becomes more important to make rapid and correct decisions.

Big data management takes data from different sources and analyzes them to find answers that save money and time, optimized proposal at the same time making intelligent decisions. Increment of data volume with industry 4.0, big data management, which uses data mining techniques, has gained more importance to obtain useful results. As a result, the linking of data received from different sources is an important problem. Russom asked, "What problems hinder the successful management of big data in your organization?" for his survey based on 2,287 responses from 461 respondents; 5 responses per respondent, on average. The report of the survey indicates that "Data integration complexity" ranked 4th with 30%. First three problems are Inadequate staffing or skills (40%), Lack of governance or stewardship (33%), Lack of business sponsorship (33%) (Russom, 2013).

Data linkage is the process of finding similar records by matching fields in different data sources, such as data files, books, internet pages, or databases. It can be named differently in various application areas. Whereas it is called list washing or merge/purge processing in mail and database applications, in computer applications, it is named a data matching or object identity problem. Data linkage is also known as coreference/entity/identity/name/record resolution, entity disambiguation/linking, duplicate detection, deduplication, record matching, reference reconciliation, object identification, data/information integration, conflation (Singla & Domingos, 2006).

Preparing and clearing data is one of the first things that need to be done in the data mining. In order to apply the methods to be used and obtain good results, the data must be standardized. It is often important to address the full data sets that are found in different data sources to achieve the right result. There are some challenges as well as the advantages of using different data sources. Data linkage is a good way to overcome these challenges (Bloomrosen & Detmer, 2008; Lipscomb, Gotay, & Snyder, 2005; Brookhart, Glynn, Rassen, & Schneeweiss, 2010; Michelle, 2014).

Halbert Louis Dunn (1946) did the first work of data linkage. In 1959, Howard Borden Nowcombe and his friends (1959) extended the data linkage theory by examining the probabilistic states. Ivan Fellegi and Alan Sunter (1969) proved that comparative data labels yield optimal results when they are independent. The