Chapter 9
A Brief Study of Approaches to Text Feature Selection

Ravindra Babu Tallamaraju
Flipkart Internet Private Limited, India

Manas Kirti
Flipkart Internet Private Limited, India

ABSTRACT

With reducing cost of storage devices, increasing amounts of data is being stored and processed for extracting intelligence. Classification and clustering have been two major approaches in generating data abstraction. Over the last few years, text data is dominating the types of data shared and stored. Some of the sources of such datasets are mobile data, e-commerce, and wide-range of continuously expanding social-networking services. Within each of these sources, the nature of data differs drastically from formal language text to Twitter or SMS slangs thereby leading to the need for different ways of processing the data for making meaningful summarization. Such summaries could effectively be used for business advantage. Processing of such data requires identifying appropriate set of features both for efficiency and effectiveness. In the current Chapter, we propose to discuss approaches to text feature selection and make a comparative study.

1. INTRODUCTION

With the increasing ability to collect, store, and share data, there is a need to find efficient algorithms that provide insights on the data that have potential utility to business advantage in commercial organizations. Some contributors to these

DOI: 10.4018/978-1-5225-2805-0.ch009

datasets are related to mobile, social networking, content-sharing, search-engines, and e-commerce enterprises.

Data Mining algorithms help to generate abstractions from these datasets. Many definitions are in use for Data Mining. The generally accepted definition (Shapiro & Frawley, 1991; Chen, et al., 1996; Leskovec et al., 2014) is that *the data mining is a process of extraction of potentially useful and hitherto unknown information which is non-trivial, and hidden in the data.* Largeness of the dataset can broadly be characterized by the limitation of in-memory processing. With increasing volumes of datasets, large datasets are termed *massive datasets* or *big data*. Big data (Douglas, 2011) is defined by three V's: *volume*, *velocity* and *variety* of the data. In the current chapter, we refer to big data and large data equivalently. Some experts suggest additional V representing *value*.

Approaches to generating abstraction from large data as discussed in the works of Leskovec et al. (2014) and Babu et al(2013) are the following.

- Real time as well as batch processing through single, distributed or parallel computation depending on requirement and nature of resources
- Divide and conquer approaches in terms of patterns or features or both
- Operate in the compressed data domain

An important aspect while processing big data is the possibility of seeing random occurrences as meeting an experimental hypothesis. Benferroni principle as discussed in the work of Leskovec et al. (2014) suggests that if the expected number of occurrences when the data is assumed random is higher than actual number of real occurrences, the results are unlikely to be correct. Following are some of the salient characteristics of algorithms for big data abstraction.

- **Scalability:** Scalability (Bondi, 2000) refers to ability of a system to process growing volumes of data gracefully. The work formally defines attributes for a scalable system.
- **Limited Number of Database Scans:** Ability to generate an abstraction with a single or least number of database scans is necessary since multiple views of databases, like that of iterative algorithms, would become expensive in terms of computation time.
- **Data Reduction:** Working on the representative patterns of dataset, instead of working on the entire dataset. Prototypes can be generated using efficient unsupervised learning algorithms.