Chapter 6
Knowledge Discovery From Massive Data Streams

Sushil Kumar Narang
SAS Institute of IT and Research, India

Sushil Kumar
IIT Roorkee, India

Vishal Verma
MLN College, India

ABSTRACT

T.S. Eliot once wrote some beautiful poetic lines including one "Where is the knowledge we have lost in information?". Can't say that T.S. Eliot could have anticipated today's scenario which is emerging from his poetic lines. Data in present scenario is a profuse resource in many circumstances and is piling-up and many technical leaders are finding themselves drowning in data. Through this big stream of data there is a vast flood of information coming out and seemingly crossing manageable boundaries. As Information is a necessary channel for educing and constructing knowledge, one can assume the importance of generating new and comprehensive knowledge discovery tools and techniques for digging this overflowing sea of information to create explicit knowledge. This chapter describes traditional as well as modern research techniques towards knowledge discovery from massive data streams. These techniques have been effectively applied not exclusively to completely structured but also to semi-structured and unstructured data. At the same time Semantic Web technologies in today's perspective require many of them to deal with all sorts of raw data.

DOI: 10.4018/978-1-5225-2483-0.ch006

1. INTRODUCTION TO KNOWLEDGE DISCOVERY

Knowledge Discovery (KD) is a concept which involves the developments of strategies and procedures for making sense out of massive data. In recent years, data have become increasingly available in substantial amounts (petabytes or zettabytes). It has numerous sources including automation of business activities (trading, mobile communication, airline reservation, or credit card usage), online activities (social media, social networking), scientific activities (experiments, simulations, and environmental sensors), biological databases (DNA/RNA/protein structures, gene expression profiles) etc. In addition, new application scenarios like weather forecasting, artificial intelligence, earth observation satellites and so forth produce terabytes of data every day. Clearly, the massive size of data ruled out any manual approach of analyzing (make sense of) collected data. If this massive data will have to be understood at all, it must be analyzed by the use of computers. Although, there are statistical procedures available for data analysis and interpretation, but this explosive growth of data requires new intelligent techniques which can astutely transform the useful data into knowledge. Knowledge discovery is the significant process of digging out meaningful patterns from huge data using automated (or semi-automated) computational tools and techniques (Devedzic, 2002; Piatetsky-Shapiro, 1996). The goals of knowledge discovery are usually identified by business domain. For instance

- Marketing agencies make use of knowledge discovery frameworks to find patterns in the way customers purchase retail items. Once they find that many individuals purchase item A along with item B, they can easily make an appropriate and potentially successful business or marketing announcement.
- Airline companies make use of knowledge discovery systems to find patterns in which their passengers fly (routes, return flights, frequency of flying to a specific destination and so forth). Based on the patterns discovered, they can give promotional offers to frequent travelers, thus attract more customers to the company.
- Banks make use of knowledge discovery frameworks to explore the database of their credits and loans. Based on the patterns discovered, they can more successfully predict the risk of approving loan to their clients, thus increasing the quality of their business decisions.

Of course, most of these goals were well existing even before knowledge discovery was conceptualized. They have been achieved by human expertise, numerical modeling and on the basis of database OLAP (online analytical processing). However, in Knowledge Discovery, these goals are achieved by applying automated (or semi-automated) computational tools and techniques to the huge amount of stored data.