Chapter 11
Pattern Recognition for Large-Scale Data Processing

Amir Basirat
Monash University, Australia

Asad I. Khan
Monash University, Australia

Heinz W. Schmidt
RMIT University, Australia

ABSTRACT

One of the main challenges for large-scale computer clouds dealing with massive real-time data is in coping with the rate at which unprocessed data is being accumulated. Transforming big data into valuable information requires a fundamental re-think of the way in which future data management models will need to be developed on the Internet. Unlike the existing relational schemes, pattern-matching approaches can analyze data in similar ways to which our brain links information. Such interactions when implemented in voluminous data clouds can assist in finding overarching relations in complex and highly distributed data sets. In this chapter, a different perspective of data recognition is considered. Rather than looking at conventional approaches, such as statistical computations and deterministic learning schemes, this chapter focuses on distributed processing approach for scalable data recognition and processing.

INTRODUCTION

Recent advancements in computing technology and data analysis have brought forward the ability to generate enormous volumes of highly-complex data which have called for a paradigm shift in the computing architecture and large scale data processing approaches. Jim Gray, a distinguished database researcher and manager of Microsoft Research's eScience group called the shift a "fourth paradigm". The first three paradigms were defined as experimental, theoretical and, more recently, computational science (Hey, Tansly & Tolle, 2009). Gray argued that the only solution to this outgrowth of big data, commonly known as data deluge, is to develop a new set of computing tools to process and analyze the data flood as the existing computer architectures are becoming

DOI: 10.4018/978-1-4666-8122-4.ch011

more incapable of dealing with data-intensive tasks over time due to their constantly growing latency gaps between multi-core CPUs and mechanical hard disks (Gray, Bell & Szalay, 2006). In fact; with an emerging interest to leverage massive amounts of data available in open sources such as the Web for solving long standing information retrieval problems; the question remains, how to effectively incorporate and efficiently exploit immense data sets. This question brings to the forefront a crucial need for high levels of scalability in the world of big data. Thus reinforcing Moore's Law of exponential increases in computing power and solid-state memory (Moore, 2000), in which it is stated that:

The complexity for minimum component costs has increased at a rate of roughly a factor of two per year... Certainly over the short term this rate can be expected to continue, if not to increase (pg. 57).

Although this was initially referred to the transistor counts within a processor, the effect of this law seems to be applicable in almost all areas of computing, including data generation and analysis. The implications of Moore's Law are quite profound as it is one of the few stable rulers we have today, in other words it's a sort of technological barometer (Malone, 1996):

It very clearly tells you that if you take the information processing power you have today and multiply by two, that will be what your competition will be doing 18 months from now. And that is where you too will have to be (pg. 6).

This outgrowth of big data has significant implications regarding the existing developments of computing applications. According to Anderson (2011), the chief editor of Wired magazine:

Sixty years ago, digital computers made information readable. Twenty years ago, the Internet made it reachable. Ten years ago, the first search engine crawlers made it a single database. Now Google and like-minded companies are sifting through the most measured age in history, treating this massive corpus as a laboratory of the human condition. They are the children of the Petabyte Age. The Petabyte Age is different because more is different. Kilobytes were stored on floppy disks. Megabytes were stored on hard disks. Terabytes were stored in disk arrays. Petabytes are stored in the cloud. As we moved along that progression, we went from the folder analogy to the file cabinet analogy to the library analogy to - well, at petabytes we ran out of organizational analogies (pg. 769).

As human beings, our brains could be viewed as large-scale distributed and interconnected networks of sensory systems and memories. Observing, recognizing and recalling what we have seen contribute to a significant portion of the activities conducted within these large-scale networks. Provided that an optimal solution is found for the scalability problem, the internet could provide the levels of interconnectivity and complexity that bear a resemblance to the human brain. Harnessing the massive potential embodied within these distributed networks of interconnected high-performance machines may provide recognition and processing capabilities for large-scale and highly-complex data.

LARGE SCALE AND BIG DATA RECOGNITION

Transforming big data into valuable information is an issue that real-world systems must grapple with. In fact, more data translates into more effective algorithms, and thus makes sense to take advantage of the enormous amounts of data that exist. In this regard; the development of powerful high-resolution data-capture instruments and sensors, in areas such as satellite and biomedical imaging, has resulted in a massive production of