MANAGING AND MINING SENSOR DATA

MANAGING AND MINING SENSOR DATA

Edited by
CHARU C. AGGARWAL
IBM T. J. Watson Research Center, Yorktown Heights, NY, USA

Kluwer Academic Publishers
Boston/Dordrecht/London

Contents

Preface

Sensor data has become pervasive in recent years because of the popularization and wider availability of sensor technology through cheaper embedded sensor devices and RFID technology. Sensors produce large volumes of data continuously over time, and this leads to numerous computational challenges. Such challenges arise both from *accuracy* and *scalability* perspectives.

The scalability challenges of sensor data analytics have reached extraordinary proportions, with the increasing proliferation of ubiquitous and embedded sensors and mobile devices, each of which can potentially generate large streams of data. Many of these devices are internet-connected. This has enabled greater possibilities for different kinds of distributed data sharing and analytics. It has been estimated that the number of internet-connected devices has exceeded the number of people on the planet since 2008. Therefore, it is foreseeable, that in the coming years, machine generated data will dominate human-generated data by orders of magnitude, and this gap is only likely to increase with time. In this context, the challenges associated with scalable and real-time management and mining of sensor data are likely to become even more significant with time.

Sensor data mining is a relatively new area, which is now reaching a certain level of maturity. In spite of this, the data analytics researchers have often remained disconnected from the networking issues which arise during data collection and processing. While the focus of this book is clearly on the data analytics side, we have taken special care to emphasize the impact of the network-specific issues on data processing.

This book discusses the key issues in the collection, modeling and processing of sensor data. The content of the book is carefully designed to cover the area of sensor data mining comprehensively. Each chapter is written as a survey by a well known researcher from the field, so as to cover this area comprehensively. Emphasis is also provided on different applications of sensor networks. A number of newer applications such as social sensing and the internet-of-things are also discussed in this book.

The book is intended for graduate students, researchers and professors. Emphasis has been placed on simplifying the material and making it more accessible. The material in the book can be helpful to both beginners on the subject and advanced researchers. At the same time, the latest topics are covered in significant detail. It is hoped that this book will provide a comprehensive overview of the research in this field. It will be a useful guide to students, researchers and practitioners.

Chapter 1

AN INTRODUCTION TO SENSOR DATA ANALYTICS

Charu C. Aggarwal
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598
charu@us.ibm.com

Abstract The increasing advances in hardware technology for sensor processing and mobile technology has resulted in greater access and availability of sensor data from a wide variety of applications. For example, the commodity mobile devices contain a wide variety of sensors such as GPS, accelerometers, and other kinds of data. Many other kinds of technology such as RFID-enabled sensors also produce large volumes of data over time. This has lead to a need for principled methods for efficient sensor data processing. This chapter will provide an overview of the challenges of sensor data analytics and the different areas of research in this context. We will also present the organization of the chapters in this book in this context.

Keywords: Sensor data, stream processing

1. Introduction

Recent years have seen tremendous advances in hardware technology such as the development of miniaturized sensors, GPS-enabled devices, pedometers, and accelerometers, which can be used to collect different kinds of data [6]. This has lead to a deluge of tremendous amounts of real-time data, which can be mined for a variety of analytical insights. The costs of sensor hardware has been consistently going down over the past few years. Furthermore, many data collection technologies [5] such as RFID have been enabled in a very cost-effective way, as a result of which the *scale* of the collection process has become enormous. Sensor data is produced in the context of a wide variety of applications such as the following:

- A wide variety of mobile devices are now GPS-enabled. This has lead to unprecedented opportunities in the context of several applications such as social sensing [4]. GPS data is also available in the context of many location-aware devices and applications.

- The decreasing cost of RFID tags has lead to tremendous volumes of RFID data. The cost of an RFID tag is now in the range of under 5 cents. This has allowed cost-effective deployment of RFID tags on products of even modest price. RFID data poses numerous challenges because of the tremendous amounts of noise in the collected data [5].

- Numerous military applications use a wide variety of sensors in order to track for unusual events or activity. This could include visual or audio cameras, or seismometers for tracking movements of large objects [9].

- Sensors are also deployed in the context of a wide variety of environmental applications, such as detecting weather and climate trends [7], and tracking pollution levels in water networks [11].

Sensor data brings numerous challenges with it in the context of data collection, storage and processing. This is because sensor data processing often requires efficient and real-time processing from massive volumes of possibly uncertain data. Some of these challenges may be enumerated as follows:

- Data collection is a huge challenge in the context of sensor processing because of the natural errors and incompleteness in the collection process. Sensors often have limited battery life, because of which many of the sensors in a network may not be able to collect or transmit their data over large periods of time. The errors in the underlying data may lead to uncertainty of the data representation [8]. Therefore, methods need to be designed to process the data in the presence of uncertainty.

- Sensors are often designed for applications which require *real-time processing*. This requires the design of efficient methods for stream processing [1]. Such algorithms need to be executed in one pass of the data, since it is typically not often possible to store the entire data set because of storage and other constraints.

- The large volumes of data lead to huge challenges in terms of storage and processing of the data. It has been estimated that since 2008, the number of internet-connected devices has exceeded the number of people on the planet. Thus, it is clear that the amount of machine generated data today greatly exceeds the amount of human generated data, and this gap is only likely to increase in the forseeable future. This is widely known as the *big data* problem in the context of analytical applications [10], or the *information overload problem* in stream processing.

- In many cases, it is critical to perform *in-network processing*, wherein the data is processed within the network itself, rather than at a centralized service. This needs effective design of distributed processing algorithms, wherein queries and other mining algorithm can be processed within the network in real time [12].

In this book, we will provide an overview of the key areas of research in sensor processing, as they related to these challenges. We will also study a number of new applications of sensor data such as social sensing, mobile data processing, RFID processing, and the internet of things.

This chapter is organized as follows. In the next section, we will discuss the key areas of research in sensor processing, as they relate to the afore-mentioned challenges. We will also relate the different research areas to these challenges. Section 3 discusses the conclusions and summary.

2. Research in Sensor Processing

The research issues in the area of sensor processing arise along all stages of the pipeline, beginning from data collection, cleaning, data management, and knowledge discovery and mining. Furthermore, many research issues arise in the context of in-network processing, which are specific to the particular application domain. The specificity to the application domain may arise in the context of other parts of the pipeline as well. Therefore, we summarize the key research issues which arise in the context of sensor data processing as follows:

- **Data Collection and Cleaning Issues:** Numerous issues arise in the context of collection of sensor data. Sensor data is inherently noisy and uncertain, and may either have missed readings or redundant readings depending upon the application domain. For example, in the context of RFID data, almost 30% of the readings are dropped, and multiple sensors may track the same RFID object. In the context of battery-driven sensors, numerous errors may arise during data transmission, and there may also be significant incompleteness because of limited battery life.

- **Data Management Issues:** The large volumes of collected data poses significant challenges for the collected data. Sometimes, the volume of the data is so large, that it may be impractical to store the entire raw data, and it may be desirable to either compress or drop portions of the data. What parts of the data should be dropped or compressed? The errors and uncertainty in sensor data, have spurred the development of algorithms for uncertain database management [2].

- **Sensor Data Mining and Processing:** The large volumes of sensor data necessitate the design of efficient one-pass algorithms which require at most one scan of the data. These are traditionally referred to as data stream mining algorithms. Furthermore, it may sometimes be advantageous to perform *in-network processing*, which can perform partial processing of the data in the network before sending these results on to a higher level of storage.

- **Application-Specific Issues:** Sensor data can arise in many domains such as retail data (RFID), military sensor networks, astronomy, the environment, and mobile data. Different domains may lead to different issues in the context of storage and processing. For example, RFID data may have larger levels of redundancy and uncertainty, whereas mobile data mining applications may require spatio-temporal mining techniques.

The different chapters of this book will study these different aspects of sensor stream processing. Therefore, the book will be organized so as to comprehensively study these different aspects. The different topics covered by the chapters of this book are as follows:

Data Collection and Management Issues The key data collection and management issues are discussed in Chapter 2. This chapter discusses some of the key database management aspects, which have recently been designed in the context of sensor data. Issues involving data uncertainty and query processing are discussed in this chapter, especially in the context of sensor data. The area of indexing and query processing is very important in the context of sensor data, and therefore we have also designed chapters specifically for this topic.

Query Processing of Sensor Data Sensor data poses numerous challenges from the perspective of indexing and query processing, because of the massive volume of the data which is received over time. A special case of query processing in sensor data is that of event detection, wherein continuous queries are posed on the sensor data in order to detect the underlying events. The main challenge in event processing is that the high level semantic events are often a complex function of the underlying raw sensor data. In some cases, the event-query cannot be posed exactly, since the event detection process is ambiguously related to the underlying data. Methods for query processing of sensor data are discussed in Chapter 3. Specialized methods for event processing of sensor data are discussed in Chapter 4.

Mining Sensor Data A variety of data mining methods such as clustering, classification, frequent pattern mining, and outlier detection are often applied to sensor data in order to extract actionable insights. This data usually needs to be compressed and filtered for more effective mining and analysis. The main challenge is that conventional mining algorithms are often not designed for real time processing of the data. Therefore, new algorithms for sensor data stream processing need to perform the analytics in a single pass in real time. In addition, the sensor scenario may often require *in-network* processing, wherein the data is processed to higher level representations before further processing. This reduces the transmission costs, and the data overload from a storage perspective. The problems of stream compression [3] and stream mining are therefore tightly integrated together from an efficiency perspective. For example, compression and hidden variable modeling provides summarized representations which can be leveraged for applications such as forecasting and outlier analysis. A survey of methods for dimensionality reduction, compression and filtering of sensor streams is provided in Chapter 5. This chapter studies the issue of stream correlation analysis, compression across streams in terms of hidden variables, and compression across time in a given stream. The application of these concepts to a few stream mining problems is also studied in the same chapter. A number of methods for real-time sensor stream mining, processing and analytics are discussed in Chapters 6 and 7. Specific methods for mining sensor streams in the distributed setting are presented in Chapter 8.

Social Sensing Applications and Mobile Data The popularity of mobile phones and other sensor-enabled devices has lead to a plethora of "socially-aware data" which can be mined in the context of a wide variety of applications. This trend has lead to the integration of sensors and dynamic social networks. A number of architectural, privacy and trust issues arise in the collection of socially aware sensor data. These issues are discussed in detail in Chapter 9. The chapter also discusses the issues of mining the different kinds of GPS- and content-based data generated in such applications.

Much of the data in social sensing applications often contains GPS trajectory data. Mobile data has a number of characteristics, which can be exploited in order to create more efficient methods for clustering, classification, anomaly detection, and pattern mining. Therefore, we have included a chapter which discusses algorithms for mobile data analysis in detail. Chapter 10 provides a detailed discussion of a wide variety of indexing and mining algorithms in the context of mobile data.

RFID Data and the Internet of Things The trend towards ubiquitous and embedded sensing has lead to a natural focus on machine-to-machine (M2M) paradigms in sensor processing. These paradigms use small RFID sensors to collect data about many smart objects. The data generated from such applications can be shared by different devices for heterogeneous fusion and inference, especially if the devices are connected to the internet. A number of issues also arise about how such devices can be effectively discovered and used by different network participants. Chapter 11 provides an overview on RFID applications for collecting such data. Issues about how such data can be used in the context of the internet of things are discussed in Chapter 12.

Software Bug Tracing in Sensor Networks Most of the aforementioned chapters provide application-specific insights on the basis of the collected data. Sensors also produce diagnostic data, which can be used in order to determine diagnostic bugs within the sensor software. Thus, this kind of mining process can be used in order to improve the performance of the underlying sensor network. A survey of methods and algorithms for software bug tracing in sensor networks is provided in Chapter 13.

Healthcare Applications Sensor data has found increasing application in the health care domain. A wide variety of Intensive Care Unit (ICU) applications use sensors such as ECG, EEG, blood pressure monitors, respiratory monitors, and a wide variety of other sensors in order to track the condition of the patient. The volume of such data is extremely large and the inferences from such data need to be performed in a time-critical fashion. Chapter 14 provides an overview of sensor mining applications in the context of health-care data.

Environmental and Climate Applications A wide variety of sensors are used in order to track environmental and sensor data. A tremendous amount of sensor data is available through satellite sensing, and other more conventional forms of sensing. Such data can be used in order to determine the short terms and long terms trends in climate change, and other environmental applications, such as detecting changes in land cover. Chapter 15 provides an overview of how sensor data may be used in the context of environmental and climate applications.

3. Conclusions and Summary

In this chapter, we provided an overview of the challenges and the key areas of research in sensor processing. We also presented the organization of this book, as it relates to these challenges. The ubiquity and volume of sensor data is likely to increase over time, as more and more applications containing sensor data become widely available. A number of emerging areas of research such as social sensing have brought the use of sensor data within the reach of the masses, because of their incorporation in commoditized devices such as mobile phones. Furthermore, newer applications such as the internet of things have lead to a greater focus on the effective storage and processing of sensor data. This book will discuss all of these challenges in a holistic and integrated way.

References

[1] C. C. Aggarwal. Data Streams: Models and Algorithms, *Springer*, 2007.

[2] C. C. Aggarwal. Managing and Mining Uncertain Data, *Springer*, 2009.

[3] C. C. Aggarwal, P. S. Yu. A Survey of Synopsis Construction Algorithms in Data Streams, *Data Streams: Models and Algorthms*, Springer, 2007.

[4] C. C. Aggarwal, T. Abdelzaher. Integrating Sensors and Social Networks, *Social Network Data Analytics, Springer*, 2011.

[5] C. C. Aggarwal, J. Han. A Survey of RFID Data Processing, *Managing and Mining Sensor Data*, Springer, 2013.

[6] I. Akylldiz, W. Su, Y. Sankarasubramaniam, E. Cayirci. A Survey on Sensor Networks, *IEEE Communications Magazine*, 2002.

[7] D. Culler, D. Estrin. M. Srivastava. Guest Editor's Introduction: An Overview of Sensor Networks. *Computer*, 37(8), 2004.

[8] A. Deshpande, C. Guestrin, S. Madden, J. Hellerstein, W. Hong. Model-driven data acquisition in sensor networks, *VLDB Conference*, 2004.

[9] D. Hall, J. Llinas. An Introduction to Multi-sensor data fusion, *Proceedings of the IEEE*, 85(1), 1997.

[10] S. Lohr. The age of big data, *New York Time Sunday Review*, February 12, 2012. `http://www.nytimes.com/2012/02/12--/sunday-review/big-datas-impact-in-the-world.html?pagewanted=all`

[11] A. Ostfeld(et.al.) The battle of the water sensor networks (BWSN): A design challenge for engineers and algorithms. *Journal of Water Resources Planning and Management*, vol. 134, no. 6, pp. 556–568, 2008. [Online]. Available: `http://link.aip.org/link/?QWR/134/556/1`

[12] Y. Yao, J. E. Gehrke. The Cougar Approach to In-Network Query Processing in Sensor Networks. *SIGMOD Record*, Vol. 31(3), September 2002.