Assuring Data Trustworthiness - Concepts and Research Challenges[*]

Elisa Bertino and Hyo-Sang Lim

Department of Computer Science, Purdue University, USA
{bertino,hslim}@cs.purdue.edu

Abstract. Today, more than ever, there is a critical need to share data within and across organizations so that analysts and decision makers can analyze and mine the data, and make effective decisions. However, in order for analysts and decision makers to produce accurate analysis and make effective decisions and take actions, data must be trustworthy. Therefore, it is critical that data trustworthiness issues, which also include data quality, provenance and lineage, be investigated for organizational data sharing, situation assessment, multi-sensor data integration and numerous other functions to support decision makers and analysts. The problem of providing trustworthy data to users is an inherently difficult problem that requires articulated solutions combining different methods and techniques. In the paper we first elaborate on the data trustworthiness challenge and discuss a framework to address this challenge. We then present an initial approach for assess the trustworthiness of streaming data and discuss open research directions.

Keywords: Data Trustworthiness, Data Integrity and Quality, Security, Policy.

1 Introduction

Today, more than ever, there is a critical need to share data among organizations so that analysts and decision makers can analyze and mine the data, and make effective decisions. Meanwhile, in many recent applications such as traffic monitoring systems and power grid management systems, a large amount of data that can convey important information for critical decision making is collected from distributed sources. In order for analysts and decision makers to produce accurate analysis, make effective decisions, and take actions, data must be trustworthy. Therefore, it is critical that data trustworthiness issues, which also include data quality and provenance, be investigated for organizational data sharing, situation assessment, multi-sensor data integration, and numerous other functions to support decision makers and analysts. Without trustworthiness, the

[*] The authors have been partially supported by AFOSR grant FA9550-07-1-0041 "Systematic Control and Management of Data Integrity, Quality and Provenance for Command and Control Applications".

W. Jonker and M. Petković (Eds.): SDM 2010, LNCS 6358, pp. 1–12, 2010.

usefulness of data becomes diminished as any information extracted from them cannot be trusted with sufficient confidence.

It is thus important to provide a comprehensive solution for assessing and assuring the trustworthiness of the information collected in such data sharing systems since decisions and analyses are largely affected by this information. Attacks or unexpected accidents may result in bad data being provided to critical components of the system. These components may in turn take wrong decisions or generate inaccurate analyses that can result in damages to real-world objects such as manufacturing facilities or power plants. For example, in an electric power grid which consists of 270 utilities using a SCADA (Supervisory Control and Data Acquisition) system that can contain up to 50,000 data collection points and over 3,000 public/private electric utilities, any single point of failure can disrupt the entire process flow and can potentially cause a domino effect that shuts down the entire systems [11].

In general the problem of providing "good" data to users and applications is an inherently difficult problem which often depends on the application and data semantics as well as on the current context and situation. In many cases, it is crucial to provide users and applications not only with the needed data, but with also an evaluation indicating how much the data can be trusted. Being able to do so is particularly challenging especially when large amounts of data are generated and continuously transmitted across the system. Also solutions for improving the data, like those found in data quality, may be very expensive and may require access to data sources which may have access restrictions, because of data sensitivity. Also even when one adopts methodologies to assure that the data is of good quality, errors may still be introduced and low quality data be used; therefore, it is important to assess the damage resulting from the use of such data, to track and contain the spread of errors, and to recover. The many challenges of assuring data trustworthiness require articulated solutions combining different approaches and techniques including data integrity, data quality, and data provenance. In this paper we discuss some components of such a solution and highlight relevant research challenges.

Our approach for assuring information trustworthiness is based on a comprehensive framework composed of two key elements. The first element is represented by trust scores that are to associated with all data items to indicate the trustworthiness of each data item. Trust scores can be used for data comparison or ranking. They can be also used together with other factors (e.g., information about contexts and situations, past data history) for deciding about the use of the data items. Our framework provides a trust score computation method which is based on the concept of data provenance, as provenance gives important evidence about the origin of the data, that is, where and how the data is generated. The second element of our framework is the notion of *confidence policy* [8]. Such a policy specifies a range of trust scores that a data item, or set of data items, must have for use by the application or task. It is important to notice that the required range of trust scores depends on the purpose for which the data have to be used. In our framework, confidence policies are integrated