Data Provenance

Vikram Sorathia
Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), India

Anutosh Maitra
Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), India

INTRODUCTION

In recent years, our sensing capability has increased manifold. The developments in sensor technology, telecommunication, computer networking and distributed computing domain have created strong grounds for building sensor networks that are now reaching global scales (Balazinska et al., 2007). As data sources are increasing, the task of processing and analysis has gone beyond the capabilities of conventional desktop data processing tools. For quite a long time, data was assumed to be available on the single user-desktop; and handling, processing as well as analysis was carried out single-handedly. With proliferation of streaming data-sources and near real-time applications, it has become important to make provisions of automated identification and attribution of data-sets derived from such diverse sources. Considering the sharing and re-use of such diverse data-sets, the information about: the source of data, ownership, time-stamps, accuracy related details, processes and transformations subjected to it etc. have become essential. The piece of data that provide such information about the given data-set is known as *Metadata*.

The need is recognized for creating and handling of metadata as an integrated part of large-scale systems. Considering the information requirements of scientific and research community, the efforts towards the building global data commons have came into existence (Onsrud & Campbell, 2007). A special type of service is required that can address the issues like: explication of licensing & Intellectual Property Rights, standards based automated generation of metadata, data provenance, archival and peer-review. While each of these terms is being addressed as individual research topics, the present article is focused only on Data Provenance.

BACKGROUND

In present scenario, data-sets are created, processed and shared at ever increasing quantities sometimes approaching gigabyte or terabyte scales. The given data-set is merely a useless series of characters, unless proper information is provided about it's content and quality. A good analogy provided to understand the importance of such metadata; a data-set can be compared with a piece of art. The art collectors and scholars are able to appreciate given object only based on authentic documented history that reveals the information like the artist, the era of creation, related events and any special treatments it was subjected to. Similarly, a given data-set may be a raw and uncorrected sensor log, or it can be derived after subjecting it to a series of careful statistical and analytical operations, making it useful for decision making.

Any kind of information provided about the data-set is therefore helpful in determining its true value for potential users. But a general record of information about given data-set; which also qualifies to be a Metadata, renders only a limited value to the user by providing simple housekeeping and ownership related information. It is only through a systematic documented history about source, intermediate stages and processing subjected to the given data-set captured in the metadata content, that provides due recognition of the quality and characteristic of the given data-set. This special type of metadata content is known as "Data Provenance" or "Pedigree" or "Lineage" or "Audit Trail". It is now being recognized (Jagadish et al., 2007) that apart from the data models, Data Provenance has critical role in improving usability of the data.

MAIN FOCUS

Provenance is used in art, science, technology and many other fields for a long time. With the recent developments of database management systems (DBMS) responding to increasingly complex utilization scenarios, the importance of data provenance has become evident. In simplest from, data provenance in DBMS can be utilized to hold information about how, when any why the Views are created. Sensor networks, enterprise systems and collaborative applications involve more complex data provenance approaches required to handle large data stores (Ledlie, Ng, & Holland, 2005).

The purpose of this article is to introduce only the important terms and approaches involved in Data Provenance to the Data Mining and Data Warehousing community. For the detailed account of the subject including historical development, taxonomy of approaches, recent research and application trends, the readers are advised to refer (Bose & Frew, 2005) and (Simmhan, Plale, & Gannon, 2005).

Characteristics of Provenance

Data Provenance is a general umbrella term under which many flavors can be identified based on how it is created and what specific information it provides. In a simple way (Buneman, Khanna, & Tan, 2001) it is classified it in two terms: *Where Provenance* and *Why provenance*. Where provenance refers to the data that provides information about location at which the source of data can be found. Why provenance provide information about the reason due to which the data is in current state. Later, (Braun, Garfinkel, Holland, Muniswamy-Reddy, & Seltzer, 2006) identified more flavors of data provenance based on the handling approach. A manual entry of provenance information is identified as *Manual Provenance* or Annotation. The approach that relies on the observation by the system is called *Observed Provenance*. In case of observed provenance, the observing system may not be configured to record all possible intermediate stages that the given dataset may pass through. Hence, resulting provenance, holding only partial information due to incomplete information flow analysis is classified as *False Provenanc*e. When participating entities provide explicit provenance to the third-party provenance handling system, it is known as *Disclosed Provenance*. When a participating system directs the desired structure and content about the object, it is classified as *Specified Provenance*. Many other provenance characteristics are studied for the detail classification of data provenance in (Simmhan, Plale, & Gannon, 2005). It mentions the reference to "*Phantom Linage"*-a special status of provenance when the data-set that is being described in given provenance record is deleted from the source.

Provenance System Requirements

Comprehensive application scenario followed by detailed use cases analysis has been documented to reveal the requirements for building provenance systems (Miles, Groth, Branco, & Moreau, 2007). It is recommended that a system targeted to handle data provenance, must support some general features apart from any specific that can be identified based on the application domain. A core functional requirement is the activity of *Provenance Record* that requires the recording the provenance using manual or automated mechanism. The collected record must be stored and shared for the utilization - which leads to the requirement for a *Storage Service*. The recorded provenance is accessed by the end-users or systems by issuing appropriately formed queries. The support for query is an important feature that allows retrieval of provenance. Handling provenance is an event-driven methodology for tracking, executing and reporting series of execution steps or processes collected as a workflow. To support the execution of a workflow enactment service is required. A processing component is required that is responsible for creation and handling, validating and inference of provenance data. Any special Domain Specific information that is beyond the scope of general purpose provenance system, and must be provided with special provisions. *Actor side recording* is a functional requirement that allows recording of provenance data from actor side that cannot be observed or recorded by an automated system. Presentation functionality is required that can allow rendering of search-results in user-friendly manner that reveals the historical process hierarchy and source information.

Some Provenance Systems

Over a period of time, many provenance systems are proposed that realize one or more functionalities identified above. In Data Warehousing and Data Mining domain, it is conventional that the users themselves assume the responsibility for collecting, encoding and handling of the required data-sets from multiple sources and maintenance the provenance information. Such manually curated data provenance poses a challenge