Understanding and Modeling Context in Data Integration

William T. Sabados, The University of Alabama in Huntsville, Huntsville, AL, USA
Harry S. Delugach, The University of Alabama in Huntsville, Huntsville, AL, USA

ABSTRACT

The pragmatic context of information is a fundamental characteristic that is not often formally addressed in data integration. This paper discusses the challenges of modeling the multiple contexts at play in data integration. A simple data integration context modeling framework is introduced that we believe addresses important issues of representing a pragmatic context. It allows for multiple data sources from similar domains to be brought together without having to designate one as the “true” semantics. An example is provided showing how this approach supports integration efforts.

Keywords: Conceptual Graph, Data integration, Multiple Interpretations, Pragmatic Context, Semiotics

INTRODUCTION

Context is a fundamental characteristic of information that is not often formally addressed in data integration. Whenever data is modeled and/or captured there is a set of assumptions and decisions made that underlie the data being modeled and/or captured. These assumptions and decisions form the context of the data and are often not well documented (if they are documented at all). The original context lives in the mind of the (human) data originator, who is often not the same individual who is attempting to integrate the data source. During integration, an analyst will examine each data source and attempt to reverse engineer the context of the data in order to properly integrate the data. Once the context is reverse engineered in this way, it is likely to be forgotten again unless extra effort is made to document it. A framework we call a conceptual graph driven knowledge space provides a mechanism to capture context once it has been discovered so it can be saved for future use. This sets up an iterative mechanism of semantic improvement for data integration. The conjecture is that stakeholders will be willing to invest extra effort to document the context of data sources that are important for a particular integration purpose and may forego the extra effort for data sources that are not as important. This paper sets out to explore context as a first class data integration concern and develops a framework to capture and improve context information for integration.

DOI: 10.4018/ijcssa.2014010101
OBSTACLES TO CAPTURING CONTEXT

If context is important to data integration, why is it not already captured and considered in current approaches? There are a number of contributing factors that may help explain why context is not generally considered a first class concern in terms of data integration. First, many data sources are not created with the intent of data integration in mind. The original goal in creating the data source may possibly be for a very specific use. Data integration may just be an opportunity to leverage existing captured data. When originally capturing the data, however, much of the knowledge forming the context of the data may simply be implicit information. For instance, if the goal is to capture information about students at a university, then the data source may neglect to document which university the students are attending because the identity of the university is understood.

Another reason implicit or assumed information may not be captured in a data source is simply because it is not efficient to do so. Data sources are primarily designed to be fast and efficient data storage and retrieval solutions. One way to build more efficient data sources is to remove or reduce redundant information, which in turn may begin to strip out some of the information that might be useful for determining context. Also, when designing a new data source, one may question the value of including information that is consistently the same value for every tuple in the data source (such as the university in the above example). If included, information that is non-variant for every tuple in the database, while useful for establishing context, will result in wasted storage space. For instance, is it worthwhile indicating that every student in a database goes to a particular university if the database is being designed solely to track students at a single university? This information is either never considered for inclusion in the data source because it is too obvious (implicit) or factored out of the data source because it is inefficient to store. Ideally this information should be added to the data source metadata in order to provide a context of what the data source designers intended to capture and store.

A third reason for not recording the context of a data source is simply that the people who created the database expect that they themselves will be performing whatever data integration is necessary in the future. The implicit context information is so “obvious” to them (having designed and constructed the data source) that they may not think any context information is worth recording (or at least, not deserving of much time and effort).

Capturing data source metadata is hardly a new or novel idea. Many data source solutions provide some capabilities to annotate data sources with designer comments. However, data integration remains a difficult task and few if any systems make a pointed effort to establish a data source context. Perhaps the chore of integration would be made easier if context information were more readily available.

PREVIOUS APPROACHES TO CAPTURING CONTEXT

The importance of context has been acknowledged before and a number of modeling implementations exist in the literature (Bettini et al., 2010; Strang & Linhoff-popien, 2004). Indeed, the realization that context is important in understanding and using information is not even particularly recent (Ross, 1964). Early investigations, such as (Ross, 1964) were largely confined to the field of natural language understanding. More recently renewed interest in storing and processing context was kindled by mobile and pervasive computing (Najar, Saidani, Kirsch-Pinheiro, Souveyet, & Nurcan, 2009; Rho, Appeltauer, Lerche, Cremers, & Hirschfeld, 2011; Schmidt, Beigl, & Gellersen, 1999). While both of these research areas acknowledge the value of context, they have different goals and use different definitions of context.

Natural language understanding is a long standing effort in artificial intelligence (Luger,
Ordered Intuitionistic Fuzzy Soft Sets and Its Application in Decision Making Problem
*International Journal of Fuzzy System Applications* (pp. 76-98).
www.igi-global.com/article/ordered-intuitionistic-fuzzy-soft-sets-and-its-application-in-decision-making-problem/208629?camid=4v1a