Chapter 5
Design of Semi-Structured Database System:
Conceptual Model to Logical Representation

Anirban Sarkar
National Institute of Technology, Durgapur, India

ABSTRACT

The chapter focuses on a graph – semantic based conceptual data model for semi-structured data, called Graph Object Oriented Semi-Structured Data Model (GOOSSDM), to conceptualize the different facets of such system in object oriented paradigm. The model defines a set of graph based formal constructs, varieties of relationship types with participation constraints. It is accompanied with a rich set of graphical notations and those are used to specify the conceptual level design of semi-structured database system. The approach facilitates modeling of irregular, heterogeneous, hierarchical, and non-hierarchical semi-structured data at the conceptual level. The GOOSSDM is also able to represent the mixed content in semi-structured data. Moreover, the approach is capable to model XML document at conceptual level with the facility of document-centric design, ordering and disjunction characteristic. The chapter also includes a rule based transformation mechanism for GOOSSDM schema into the equivalent XML Schema Definition (XSD). Moreover, the chapter also provides comparative study of several similar kinds of proposals for semi-structured data models based on the properties of semi-structured data and future research scope in this area.

DOI: 10.4018/978-1-4666-2958-5.ch005
INTRODUCTION

The increasingly large amount of data processing on the web based applications has led a crucial role of semi-structured database system. In recent days, semi-structured data has become prevalent with the growing demand of such internet based software systems. Semi-structured data though is organized in semantic entity but does not strictly conform to the formal structures of strict types. Rather it possesses irregular and partial organization (Abiteboul, 1999). Further semi-structured data evolve rapidly. Thus, unlike structured database system, the schema for such data is large, dynamic, is not strict to type and also is not considered the participation of instances very strictly.

The eXtensible Markup Language (XML) is increasingly finding acceptance as a standard for storing and exchanging structured and semi-structured information over internet (Conrad, 2000). The Document Type Definition (DTD) or XML Schema Definition (XSD) language can be used to define the schema which describes the syntax and structure of XML documents (Liu, 2006). However, the XML schemas provide the logical representation of the semi-structured data and it is hard to realize the semantic characteristics of such data. Thus it is important to devise a conceptual representation of semi-structured data for efficient design of the information system based on such data. For detail reference on XML technology refer W3C Standard (2008) and W3C Standard (2012) of additional reading section.

A conceptual model of semi-structured data deals with high level representation of the candidate application domain in order to capture the user ideas using rich set of semantic constructs and interrelationship thereof. Besides some similar characteristics of structured (classical) database system, several crucial characteristics are added complexity for the design of semi-structured database system. For effective design of such system, the intended conceptual model must be capable to adopt the rapidly data evolving characteristics, representation of irregular and heterogeneous structure, hierarchical relations along with the non – hierarchical relationship types, cardinality, n – array relation, ordering, representation of mixed content etc. (Necasky, 2006). Beside these, it is also important to realize the participation constraints of the instances in association with some semi-structured entity type. Even more, the participations of instances in semi-structured data model are not strict. Thus Object Oriented (OO) paradigm is most suitable to represent the organization of semi-structured data. Further, the conceptual design of semi-structured database system should be rich enough to efficiently represent such system. Such conceptual model will separate the intention of designer from the implementation and also will provide a better insight about the effective design of semi-structured data based system. The conceptual design of such system further can be implemented in XML based logical model.

With the aforementioned objectives, the chapter has been organized in six sections. In section 2 previous researches related to the semi-structured data modeling have been summarized with major emphasize on the models based on OO paradigm. In section 3, the GOOSSDM (Sarkar, 2011 June, November) has been introduced along with the prototype CASE tools and rule based transformation mechanism of conceptual model schema to its equivalent logical model schema based on XSD technology. In section 4, major characteristics of semi-structured data model have been summarized. This section also includes a comparative study of all OO semi-structured data models. In section 5, the future research directions of semi-structured database have been summarized. Finally, the chapter has been concluded in section 6.