Available online at www.sciencedirect.com
ScienceDirect
Data & Knowledge Engineering 63 (2007) 972–996
DATA & KNOWLEDGE ENGINEERING
www.elsevier.com/locate/datak

Fuzzy XML data modeling with the UML and relational data models

Z.M. Ma *, Li Yan

Department of Computer Science and Engineering, Northeastern University, Shenyang, Liaoning 110004, China

Received 21 February 2006; received in revised form 4 November 2006; accepted 7 June 2007
Available online 22 June 2007

Abstract

Information imprecision and uncertainty exist in many real-world applications and for this reason fuzzy data modeling has been extensively investigated in various data models. Currently, huge amounts of electronic data are available on the Internet, and XML has been the de facto standard of information representation and exchange over the Web. This paper focuses on fuzzy XML data modeling, which is mainly involved in the representation model of the fuzzy XML, its conceptual design, and its storage in databases. Based on ''possibility distribution theory'', we developed this fuzzy XML data model. We developed this fuzzy UML data model to design the fuzzy XML model conceptually. We investigated the formal conversions from the fuzzy UML model to the fuzzy XML model and the formal mapping from the fuzzy XML model to the fuzzy relational databases.

Keywords: Fuzzy sets and possibility distributions; XML; UML; Relational databases; Conceptual design; Mapping

1. Introduction

With the wide utilization of the Web and the availability of huge amounts of electronic data, information representation and exchange over the Web becomes important, and XML has been the de facto standard [7]. XML and related standards are technologies that allow the easy development of applications that exchange data over the Web such as e-commerce (EC) and supply chain management (SCM). This creates a new set of data management requirements involving XML, such as the need to store and query XML documents. To store, query and update XML data, it is necessary to integrate XML and databases [3]. Various databases, including relational, object-oriented, and object-relational databases, have been used for mapping to and from the XML document [10,16–19,38,45]. Among these kinds of databases, relational databases might be the more

* Corresponding author.
E-mail address: mazongmin@ise.neu.edu.cn (Z.M. Ma).

0169-023X/$ - see front matter
doi:10.1016/j.datak.2007.06.003

promising alternative because of the widespread use and mature techniques [13]. Moreover, XML lacks sufficient power in modeling real-world data and their complex inter-relationships in semantics. Thus, it is necessary to use other methods to describe data paradigms and develop a true conceptual data model, and then transform this model into an XML encoded format, which can be treated as a logical model. Conceptual data modeling of XML document schema [11,35,27,48,14] and XML schema [2,37] have been studied in the recent past. In [11], for example, UML was used for designing XML DTD (document type definition). The idea is to use essential parts of the static UML to model the XML DTD. The mapping between the static part of the UML (i.e., class diagrams) and the XML DTDs was developed. To take advantage of all facets that DTD concepts offer, the authors further extended the UML language in an UML-compliant way.

In real-world applications, however, information is often vague or ambiguous. Therefore, different kinds of imperfect information [44] have been extensively introduced into databases. Inconsistency, imprecision, vagueness, uncertainty, and ambiguity are five basic kinds of imperfect information in database systems [6].

- Inconsistency is a kind of semantic conflict, meaning the same aspect of the real world is irreconcilably represented more than once in a database or in several different databases. For example, the *age* of *George* is stored as 34 and 37 simultaneously. Information inconsistency usually comes from information integration [12].
- Of course, imprecision and vagueness are relevant to the content of an attribute value, meaning that a choice must be made from a given range (interval or set) of values without knowing which one to choose. In general, vague information is represented by linguistic values. For example, the *age* of *Michael* is a set $\{18, 19, 20, 21\}$, a piece of imprecise information, and the *age* of *John* is a linguistic "*old*", being a piece of vague information.
- The uncertainty is related to the degree of truth of its attribute value, meaning that we can apportion some, but not all, of our belief to a given value or group of values. For example, the possibility that the *age* of *Chris* is 35 right now should be 98%. The random uncertainty, described using probability theory, is not considered in the paper.
- The ambiguity means that some elements of the model lack complete semantics, leading to several possible interpretations.

Generally, several different kinds of imperfection can co-exist with respect to the same piece of information. For example, the *age* of *Michael* is a set $\{18, 19, 20, 21\}$ and their possibilities are 70%, 95%, 98%, and 85%, respectively. Also Smets [44] presents some aspects of imperfection, in which *imprecision*, *inconsistency* and *uncertainty* are the major groups. Imprecision and inconsistency are essential properties of the information itself, whereas uncertainty is a property of the relation between the information and our knowledge about the world. To model imprecision and uncertainty, the various approaches are presented in [44]. These models are grouped into two large categories, namely, *the symbolic* and *the quantitative* models. Fuzzy sets introduced by Zadeh [52] have been widely used for the quantification of imprecision and uncertainty.

Fuzzy information has been extensively investigated in the context of the relational model [8,34,36]. In order to model uncertain data and complex-valued attributes as well as complex relationships among objects, current efforts are focusing on the conceptual data models [51,26], and object-oriented databases [22], with imprecise and uncertain information. Information fuzziness has also been investigated in the context of e-commence (EC) and supply chain management (SCM) [33,50,49]. It is shown that fuzzy set theory is very useful in Web-based business intelligence. Unfortunately, although it is the current standard for data representation and exchange over the Web, XML is not able to represent and process imprecise and uncertain data, although the databases with imprecise and uncertain information have been extensively discussed. Currently, little research has been done in modeling and querying imperfect XML data. Only XML with incomplete information [1] and probabilistic data in XML [30,47] has been proposed in research papers. More recently, Lee and Fanjiang [20] developed a fuzzy OO modeling technique schema based on XML to model requirement specifications and incorporate the notion of stereotype to facilitate the modeling of imprecise requirements.

In this paper, we identify multiple granularity of data fuzziness in XML. Based on possibility distribution theory, we developed a fuzzy XML data model that addresses all types of fuzziness. Also, we developed a fuzzy UML data model to design the fuzzy XML data model conceptually. In particular, we investigated