2011 International Conference on Telecommunication Technology and Applications
Proc .of CSIT vol.5 (2011) © (2011) IACSIT Press, Singapore

X-CM: A Conceptual Modeling for XML Databases

Swee-Mei Chin†, Su-Cheng Haw and Chien-Sing Lee

Faculty of Information Technology
Multimedia University Cyberjaya, Malaysia

Abstract. Most organizations are using XML as the data exchange format over the Internet. Consequently, some are beginning to store data in the native XML database. Unfortunately, as to date, there is no suitable mechanism for intuitively modeling the components of XML conceptually. The area of designing conceptual modeling techniques for XML is still not adequately explored in literature. As such, in this paper, we proposed a novel XML conceptual modeling technique, which we named as X-CM (XML Conceptual Modeling). We explored the possibility of extending Entity-relationship modeling in the following aspects: 1) expressing the containment semantics more explicitly, and 2) specifying the data dependencies in multiple contexts.

Keywords: Conceptual modeling, XML, XML schema, Data Type Definition, XML database

1. Introduction

XML has rapidly become the de facto standard to describe data and information exchange over the World Wide Web. The increasing usage of XML documents over many application domains such as document repositories, digital libraries and business transactions had sparked the need to have a powerful schema to describe the structure of the XML instance and to define the syntax to ensure data integrity. As to date, Document Type Definition (DTD) and XML Schema Definition (XSD) are the most widely used schemas. Nevertheless, these schemas represent the logical model rather than the conceptual model. As a result, the semantics underlying these documents are difficult to express.

Conceptual modeling has a significant influence during the development phase. A person without any knowledge on the low level language is able to understand the flow of the application by looking at the graphical representation of the model. For instance, Entity-relationship (ER) modeling is very successfully used as a conceptual model in relational databases while Unified Modeling Language (UML) is widely adopted as a conceptual model in object-oriented databases. Despite the importance of conceptual modeling, XML still does not have its own conceptual model. Although many researchers proposed using ER and UML as the conceptual model for XML, unfortunately, ER and UML are not suitable to represent XML because these models lack the ability to express the semantic and special characteristics of XML such as hierarchy, ordering, schema-less content and mixed content. This motivates us to propose a novel XML conceptual modeling technique by exploring the possibility of extending the ER model.

The rest of the paper is organized as follows. Section 2 reviews the existing conceptual modeling. Section 3 presents our novel conceptual modeling technique. Finally, Section 4 concludes the paper with some discussion and future works.

2. Related Work

Conceptual modeling of XML data is an emerging area of research. Most models, which could be categorized as conceptual models, have been created in an attempt to extend or adapt the relational or object-oriented model to XML's properties. This approach of creating a modeling language has the problem that

† Corresponding author. Tel.: 03 8312 5268; fax: 03 8312 5264.
E-mail address: smchin@mmu.edu.my.

some of the main properties of XML, such as its hierarchical structure, the ability to specify alternatives, mixed content, and schema-less content, do not fit very well into the traditional database modeling world.

Feng et al. proposed a semantic network-based conceptual model for XML documents based on 2 levels, i.e., semantic level and schema level [1]. The semantics of XML such as cardinality, strong/weak adhesion and ordering are expressed through sets of nodes and edges. Different types of semantic relationship such as generalization, aggregation, association and of-property relationship were also discussed. However, this design does not support mixed content in the schema approach. Mani [2] proposed ER extended for XML (EReX) by extending the ER model with additional features such as key constraint, cardinality constraints, categories, coverage constraints, and order constraints. Categories are a type of relationship similar to is-a (ISA) relationship types. EReX support hierarchy and ordering but fail to support mixed content and other XML features. Sengupta et al. suggested another extensible ER model named XER [3]. XER supports ordering, mixed entity and generalization relationship but does not discuss supporting complex structure elements.

On the other hand, Badia proposed Extended-ER based on DTD [4]. Badia suggested marking all attributes as optional or required; and a choice attribute to add more flexibility on representing the semantic of XML documents. Nevertheless, the model still lacks the ability to express ordering and mixed content. The X-Entity [5], a conceptual modeling approach based on XSD, focuses on representation of XML structure by elements and attributes. However, the approach only supports limited XML features but not on mixed content, ordering and hierarchy structure of the document. On the other hand, ORA-SS is a powerful conceptual model which captures real world semantics [6]. The model discussed on XML specialties such as ordering on object and attribute, disjunctive and recursive relationship, functional dependency, inheritance and so on. Nevertheless, the approach did not mention mixed content features.

Some approaches are developing based on UML such as XUML [7] and UXS [8]. XUML supports most of the data types defined in W3C XML Schema definition language (WXS). Two key constructs were introduced to help increase the accuracy namely, Generic Aggregation and Business Component. The graphical component not only concentrated on the simple type, complex type and ordering of element, but considered the notation, schema documentation and inclusion of external schema. Besides, the authors developed a set of translation mechanisms to obtain XML documents with correct semantics from the UXS schema. Conrad et al. [9] proposed methods to create a mapping between UML class diagrams and XML. However, the modeling features in UML have been used to match a specific XML schema language, rather than creating an XML conceptual modeling language. Recently, Fong et al. [10] proposed XML Tree Model (XTM). XTM logical schema can be translated into XML conceptual schema through reverse engineering. These processes are done through a basic set of rule-based algorithm and an information capacity with pre- and post-conditions.

Table 1 summarizes all the features supported by each approach. From Table 1, the most basic features needed to be covered on modeling XML are object, attribute (optional, required, composite and multi-value), primary key, relationship, cardinality, ordering, connectivity and generalization (overlap, disjoint). However, there still lacks a complete design on conceptual modeling to support the entire basic feature. Thus, we proposed a new X-CM that is not only able to express the basic semantics, but is also able to handle 'any' attribute and adopt three kinds of ordering concept from ORA-SS [6].

Table 1. Summary of features supported on existing conceptual model for XML

Features	Semantic Network	ERex	Extended-ER	XER	X-Entity	ORA-SS	XUML	UXS	Conrad. et.al	XTM
Object	√	√	√	√	√	√	√	√	√	√
Attribute										
* normal	√	√	√	√	√	√		√		√
* required			√		√	√				
* optional			√		√	√				
* any						√			√	
* composite	√	√	√			√				
* multi-value			√		√	√				
* single value						√				
* fixed value						√				
* default value						√		√		
* derived value			√			√				
* primary key	√	√				√		√	√	
* candidate key						√			√	
Relationship	√	√	√	√	√	√	√			√
Cardinality	√	√	√	√	√	√		√	√	√
Ordering	√	√		√		√	√	√	√	
Connectivity			√		√	√	√			

Generalization	√	√	√	√	√	√		√	√	√
* overlap		√	√					√	√	√
*disjoint		√	√	√	√	√		√	√	√
Aggregation	√								√	
Association	√								√	
Of-property	√									
Mixed Entity				√					√	
Symmetric						√				
Generic Aggregation							√			
Business Component							√			
Reference						√				
Schema Definition								√		
Notation								√	√	
External Schema								√		
Group										√

3. X-CM Model Construction

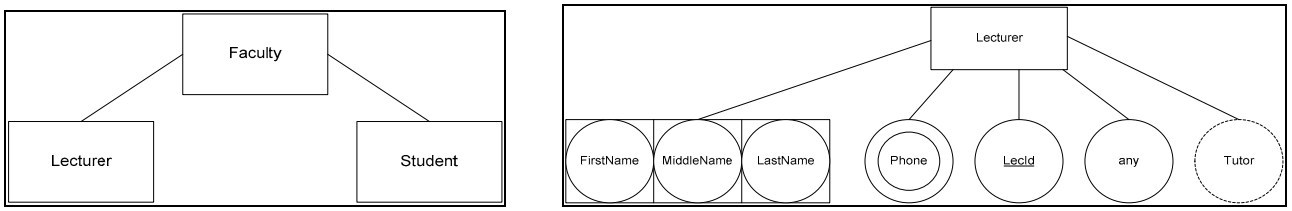
In this paper, we presented our XML conceptual model (X-CM) by extending ER data modeling (based on Chen model) [11]. Nevertheless, some modifications are done on the ER model in order to express the semantics and structure of XML documents correctly. Instead of creating a totally new conceptual model for XML, the ER model is selected as the reference because of it has been widely used in database modeling since several decades ago. The basic construct for X-CM is elaborated as below.

Entities

Entities are defined as an object in the real world and can exist either independently or dependently. An entity could be simply a person, thing, object or event from which we intend to collect the data. For example, Fig. 1 depicts the entities identified in a faculty of a University. In ER, an entity is represented as a labeled rectangle [11]. We adopted the same notation in our model as shown in Fig. 1(a).

Attributes

Attributes are used to describe the characteristic about an entity. The four main types of attributes are simple, composite, optional and multivalued. In ER construct, each attribute is represented as a labeled oval. In our proposed model, each member attribute in a composite attribute is represented by a labeled circle located inside a rectangle and is connected side by side to each other. For example, Fig. 1(b) shows the composite attribute, i.e. the lecturer's name consisting of FirstName, MiddleName and Lastname. The multivalued attribute is represented as a double layer labeled circle while the optional attribute has a dash line labeled circle. For simple attribute, we use similar notation to the ER construct. The key attributes are underlined in both ER and our model. Unlike relational database, XML has a special attribute which could appear in any form either as an entity or attribute. Therefore, we proposed a new attribute type name 'any'. 'Any' attribute is represented by using a labeled name 'any' in the circle.



(a) (b)

Fig. 1. (a) The connection between the entities Faculty, Lecturer and Student (b) The four main types of attributes and an 'any' attribute.

Relationships

Relationships are represented by a connected line and used to showed aggregations between entities and attributes. Relationships can be classified by using degree, connectivity and cardinality. Degree of a relationship is determined by identifying the numbers of entities involved in a relationship. A unary relationship (a.k.a. recursive relationship) exists if it involves only a single entity, while a binary relationship exhibits the relationship involving two entities. There are 3 basic types of connectivities, i.e., one-to-one (1:1), one-to- many (1: M) and many-to-many (M: N). One-to-one relationships means at most an entity A can associate with only 1 entity B. On the order hand, one-to-many relationships means an entity A can have association with many entities. Finally many-to-many relationship explains that many entities can association with many entities. Cardinality describes the number of occurrences of the related entities. In XML schema, the cardinality is indicated by the minOccurs and maxOccurs tag. Derived attribute can occur in a

relationship between two entities. Derived attribute is represented by a dashed line connected to a labeled circle. Fig. 2 shows an example of relationship between the entity Student and Course.

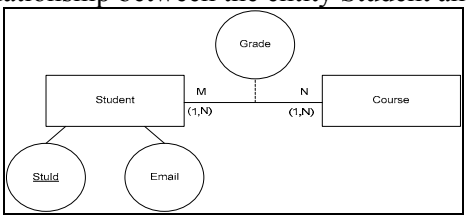
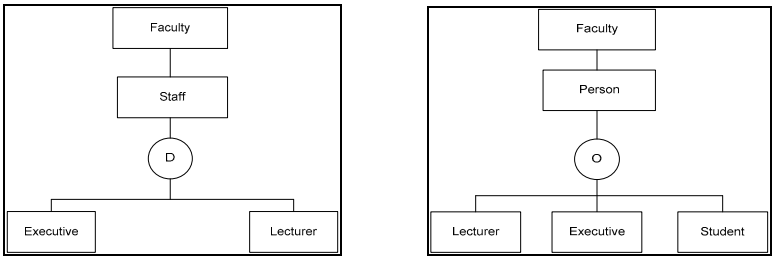


Fig. 2. The binary relationship between Student and Course; Grade is a derived attribute for a Student in a particular Course.

Generalizations

Generalizations are type of relationship similar with an ISA relationship. Generalization is a top-down approach where common attributes shared among two or more entities will be then generalized into another new entity called super entity. Thus, a new type of relationship is created whereby super entity contains subclasses or subtype. Our model has two types of generalization named disjoint generalization and overlapping generalization. Disjoint generalization means an entity can only belong to one of the subtypes. For example, faculty staff could be either an executive or a lecturer. Overlapping generalizations means an entity can belongs to many subtypes. For example, a person in a faculty can be lecturer, student and executive at the same time. Disjoint generalizations is represented by a labeled circle with alphabet 'D' (disjoint) inserted between super type and its subtypes while for overlapping generalization, the label is changed to 'O' to indicate overlapping relationship as depicted in Fig. 3.



(a) (b)

Fig. 3. (a) A Staff could be either is an Executive or a Lecturer (b) A Person can be a Lecturer, Executive and Student at the same time.

Ordering

Ordering is one of the special features found in XML documents. The *sequence* tag appearing in XML schema indicates that the element has to appear in a specific order as declared to ensure data integrity. Our model adapts the ordering concept from ORA-SS [6] model where the authors introduced three types of ordering. Yet, we refer to element in XML schemas as the object reference in ORA-SS. Below are the symbols used to indicate different types of ordering in our conceptual model.

OE : Ordering on instance of an element.

OAV : Ordering on value of an attribute.

OA : Ordering on set of attribute of an element.

OE placed beside the relationship means that all instances of the particular child elements are ordered while an OA placed beside an element indicates that all attributes under the element are ordered. On the other hand, OAV placed beside an attribute indicates that all values of the attributes are ordered as shown in Fig. 4.

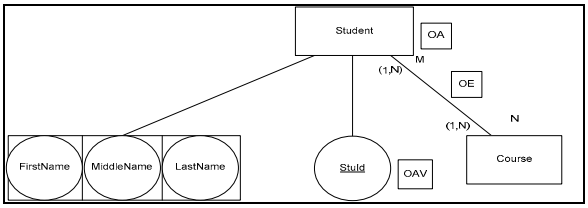


Fig. 4. Ordering representation

Fig. 5 depicts the overall proposed basic notation for the X-CM model.

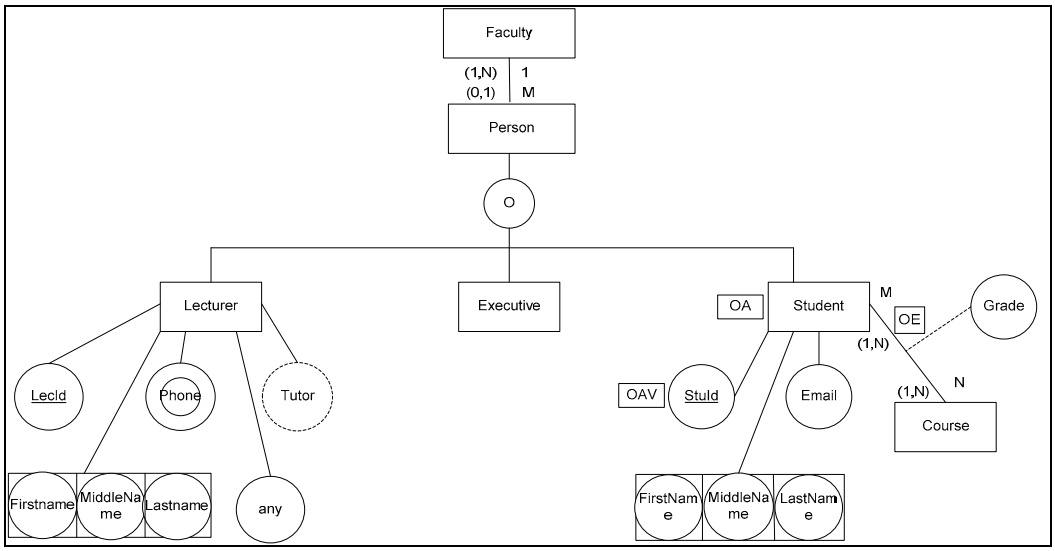


Fig. 5. Basic notation for X-CM model

4. Summary and Conclusion

Effective use of XML has caused more and more organizations to store XML in its native storage. A well-designed conceptual model for XML is crucial to produce a valuable XML document with less error and to clarify as much as possible the semantics underlying the respective XML schema. This situation has motivated us to propose a model by extending on the ER model.

Our future works include (1) to investigate some other important elements to be included in our model, (2) to validate our proposed model through case studies, and (3) to implement a framework that has the capability to automatically support up- and down- translation from conceptual modeling to scheme code generation.

5. References

[1] L. Feng, E. Chang, T. Dillon. A Semantic Network-Based Design Methodology for XML Documents. ACM Transactions on Information Systems, 20 (4), 390–421, 2002.

[2] M. Mani. EReX: A Conceptual Model for XML. In Lecture Notes in Computer Science, 3186, pp. 128–142, 2004.

[3] A. Sengupta, S. Mohan, R. Doshi. XER - Extensible Entity Relationship Modeling. In Proceedings of the XML Conference, pp. 140-154, 2003.

[4] A. Badia. Conceptual Modeling for Semistructured Data. In Proceedings of the 3rd International Conference on Web Information Systems Engineering Workshops, pp. 170-177, 2002.

[5] B.F. Losio, A.C. Salgado, L.R. Galvao. Conceptual Modeling of XML Schemas. In Proceedings of the 5th ACM International Workshop on Web Information and Data Management, 2003.

[6] G. Dobbie, X. Wu, T.W. Ling, M.L. Lee. ORA-SS: An Object-Relationship-Attribute Model for Semistructured Data, Technical Report TR21/00. National University of Singapore, 2000.

[7] H.X. Liu, Y.S. Lu, Q. Yang, XML Conceptual Modeling with XUML. In Proceedings of International Conference on Software Engineering, pp. 973 - 976, 2006.

[8] C. Combi, B. Oliboni. Conceptual modeling of XML data. Proceedings of SAC'2006, pp. 23-27, 2006.

[9] R. Conrad, D. Scheffner, J. C. Freytag. XML Conceptual Modeling Using UML. In Lecture Notes in Computer Science, 1920, pp. 558–571, 2000.

[10] J. Fong, S.K. Cheung and H. Shiu. The XML Tree Model - toward an XML conceptual schema reversed from XML Schema Definition. Data & Knowledge Engineering 64(3), pp. 624--661, 2008.

[11] P. P. Chen. The Entity-Relationship Model. ACM Transactions on Database Systems (TODS), 1:9–36, 1976.