Constraint-Based Data Transformation for Integration: An Information System Approach*

Md. Sumon Shahriar and Jixue Liu

Data and Web Engineering Lab
School of Computer and Information Science
University of South Australia, Adelaide, SA-5095, Australia
E-mail: shamy022@students.unisa.edu.au, Jixue.Liu@unisa.edu.au

Abstract

Transforming data from different information systems is important and challenging for integration purposes as data can be stored and represented in different data models in different information systems. In addition, when modeling data in the information systems, integrity constraints on the schemas are necessary for semantics and to maintain consistency purposes. Thus when schemas with conforming data are transformed from heterogeneous information systems, there is a need to transform and preserve semantics of data using constraints. Addressing this problem, we propose how data from different source information systems can be transformed to a global information system. We also review how constraints in data transformation are used in data integration for the purpose of integrating information systems. Our research is towards the handling of semantics using integrity constraints in data integration from heterogeneous information systems.

1. Introduction

As different information systems (ISs) use different data models for storing and representing data, transformation of data for exchange and integration purposes [1] is necessary and is a challenging task. Each information system is designed with its own data model and schemas are designed with constraints to convey semantics of data and for maintaining consistency. Historically most popular data model is relational data model. However, much of data is currently being represented and exchanged in XML [7] over the world wide web. Thus transformation of XML data to relational data and transformation of relational data to XML data become necessary for data intensive activities such as data integration, data warehousing, data exchange and data publishing [21, 28]. We show the data transformation and integration for heterogeneous information systems in Fig.1.

The Fig.1 shows the data transformation and integration architecture for different information systems with different data models. Both local information systems and the global information system can be either in XML or in relational database. When both local information systems and the global information system have different database systems, there is a need to transform schema

*This research is supported with Australian Research Council (ARC) Discovery Project Fund.

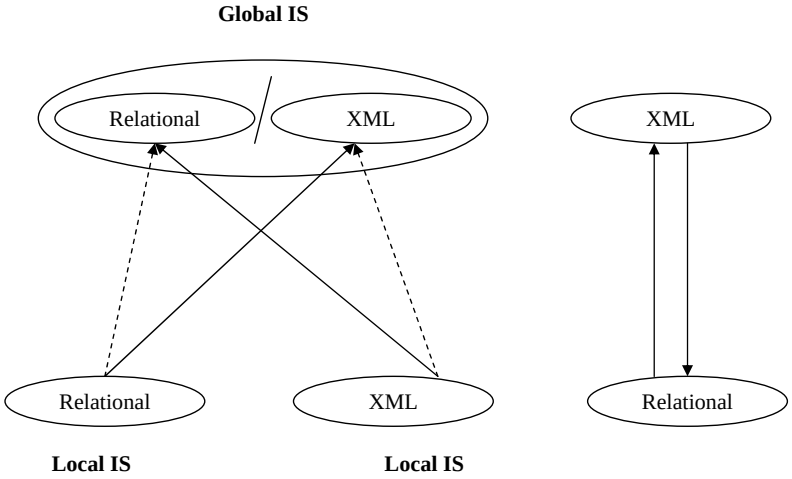

Figure 1. Data integration from heterogeneous information systems

and data with constraints. We also identify that even for the same database systems (e.g.XML) for both local and global information systems, there is a need to transform schema and data with constraints for preserving semantics.

We study the followings in this paper.

- We review the researches of data transformation and integration with constraints for different data models.
- We then identify the data transformation framework for different data models.
- A data integration framework based on different data models with constraints from different information systems is also proposed.

2. Data Transformation and Integration with Constraints: Overview

We now review the achievements of transformations of data with constraints for different data models in information systems.

2.1. Relational to XML Data Transformation [$R - X$]:

With the emergence of XML as data representation and storage format over the web, data transformation and integration from relations to XML became necessary for some purposes such as data exchange and data publishing. In relational to XML data transformation, the issue of constraints preservation is studied in [27] where the constraints like primary keys, foreign keys, unique constraints, and not null are considered when a relation is transformed to XML Schema. In publishing data in XML from XML and relational data [28], constraints are also exploited for better query formulation. In data translation [21], both XML and relational schemas are considered with some constraints like nested referential constraints.

2.2 XML to Relational Data Transformation [$X - R$]:

Like relational to XML, data transformation and integration from XML to relations also got much attention in past. In [23], constraints(e.g., cardinality constraints, domain constraints, inclusion dependencies etc.) are considered for preservation when XML DTD is transformed to relational schema. In [22], XML keys are transformed to functional dependencies(FDs) for relations and this process is termed as constraint propagation. In [25], how relational keys can be captured from XML keys is shown. Some work ([24, 26]) where constraints like cardinality constraints, inclusion dependencies, constraints on DTD etc. are considered where preservation is the issue for XML to relational data transformation perspective.

Discussions: Transformation of XML constraints are very different from relational constraints in the sense that constraints in XML follow the tree structure of data and relational constraints follow flat structure of data. Thus when transforming an XML schema with data to relational schema with data, there is a need to investigate how constraints in XML are transformed to relational constraints with preservation. Similarly, when relational schema with data is transformed to XML schema with data, there is a need to investigate how to transform constraints in relations to constraints in XML and how to preserve constraints by the transformed XML data.

We now discuss the data transformation for homogeneous data model.

2.3. Relational to Relational Data Transformation [$R - R$]:

The research works in [10, 35, 36, 37, 38] are the examples of data integration in relational database and use the transformation or restructuring of source schemas for schema integration.

Integrity constraints integration for schema integration was studied for heterogeneous databases in [10] and the research was mainly in relational data model.

McBrien and Poulovassilis [35] developed a data integration system in relational model. They use a set of primitive transformations on source schemas to integrate them into global schema. A general framework for transformation of schemas is shown in [37] by them. In [38], they described a formal framework for schema integration that uses a common data model in Entity-Relationship (ER) model. They also proposed a set of transformation operators for schema integration in [38]. They extended their research works and presented an approach named schema evolution [36] in schema transformation and integration.

In pure relational data integration [3], how queries should be affected in the presence of keys and foreign keys on the global schema and no constraints on the source schema is shown. In [15], a data integration system was shown where both source schemas and the global schema can have constraints and then he showed how the source derived global constraints and the original constraints on the global constraints can further be used for answering the query. In [16], how the consistent local constraints on the local schemas are transformed and simplified to the global schema is shown. In relational schema integration, the correspondence between local integrity constraints and the global extensional assertions is investigated in [17]. In [18], the query preserving transformation in data integration system is shown with or without constraints on the global schema. In [19], how the integrity constraints over the global schema can affect the query answering is discussed. In [20], the inconsistency of a data integration system is illustrated when source constraints over the source schemas are not the same as global constraints over the global schema.

2.4. XML to XML Data Transformation [$X - X$]:

In recent years, with the massive use of XML over world wide web, the task of data transformation and integration in pure XML is worth to mention. In [2], XML keys and foreign keys are taken into consideration on the XML global schema where source schemas are also in XML. In [30], data from relational sources are integrated to the XML target schema where keys and foreign keys in relations are captured as XML keys and XML inclusion constraints on the target schema using constraint compilation. In XML to XML data transformations and integration ([32, 14, 9, 11, 12, 13]), how the important XML constraints(e.g. XML keys, XML Functional Dependencies) on the source schema should be transformed and preserved to the target schema is investigated in [5, 6].

3. Data Transformation Framework In Heterogeneous Data Model

In data transformation for integration, a source schema with its conforming data is transformed to target schema. A source schema can be defined with integrity constraints to convey semantics of data. When a source schema is transformed to target schema with their conforming data, constraints need to be transformed and preserved. We illustrate these problems using the Fig.2.

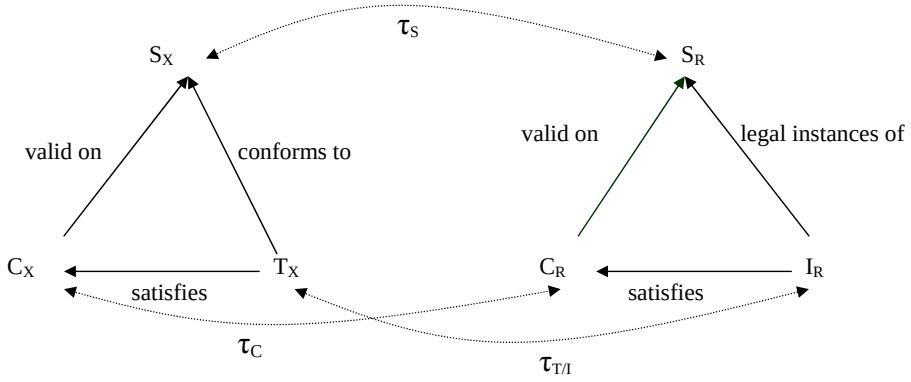

Figure 2. Data transformation of heterogeneous information systems

We denote a database of an information system as a triple $\Delta = (S, T/I, C)$ where S is schema, T/I is document or instance and C is constraint. We use $\Delta_X = (S_X, T_X, C_X)$ to mean a database in XML and $\Delta_R = (S_R, I_R, C_R)$ to mean a database in relational model.

We denote the transformation as τ that has three sub operations: the schema transformation τ_S, the document or instance transformation $\tau_{T/I}$ and the constraint τ_C.

We now study the effects of transformations on schema, document or instance and constraint.

3.1. Schema Transformation[$\tau_S(S)$]

In schema transformation, different transformation operations are used. For $\tau(\Delta_R) \rightarrow \Delta_X$, relational tables need to be transformed to different schema definitions in XML such as XML Document Type Definitions(DTD)[7] and XML Schema[8]. Similarly, for $\tau(\Delta_X) \rightarrow \Delta_R$, different schemas in XML are to be transformed to relations.

3.2. Data Transformation[$\tau_{T/I}(T/I)$]

When schemas are transformed, underlying data or documents conforming to the schemas need to be also transformed. For $\tau(\Delta_R) \rightarrow \Delta_X$, flat structure instances of relational tables are to be transformed to tree structure XML documents. Opposite task is needed with the case of $\tau(\Delta_X) \rightarrow \Delta_R$. One property known as $information\ preservation$[31] is necessary when data is transformed.

3.3. Constraints Transformation[$\tau_C(C)$]

When schemas with data are transformed, the constraints specified on the schemas need to be transformed. For $\tau(\Delta_R) \rightarrow \Delta_X$, constraints(Primary key,unique constraints,functional dependency, foreign key etc.) in relational data are to be transformed to XML constraints(XML absolute key, relative key, XML functional dependency, XML inclusion dependency etc.). When transforming relational constraints to XML constraints, the graph-structured XML schema and the tree-structured XML documents need to be considered. There is also a need to preserve the constraints [23, 24, 27] in transformations. Similarly, the opposite transformations and preservations are needed when the transformation is $\tau(\Delta_X) \rightarrow \Delta_R$.

4. Data Integration Framework with Constraints in Different Data Models for Information Systems

In data integration, from sources, schema and its conforming data with consistent constraints need to be transformed and integrated to the global site. Moreover, the global information system needs to have its own constraints for data consistency and integrity. We show this framework in the Fig.3. The global information system is denoted as I^G and the source information systems are denoted as I^{S1} and I^{S2}.

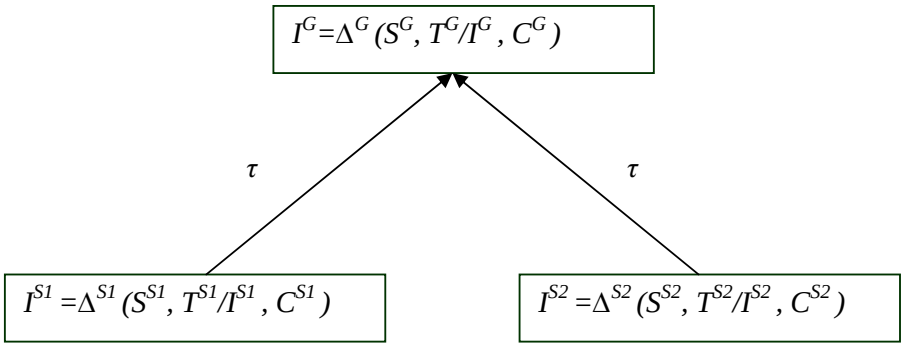

Figure 3. Data transformation and integration of heterogeneous information systems

In transforming schema S, document or instance T/I and constraint C, there is a need of a set of transformation operations. These operations need to be sufficient to transform a source schema with its conforming data and the constraints on the schema to the global schema.

We identify the following tasks need to be performed in data integration with constraints.

(a) The constraints on the source are to be transformed and preserved to the global site for semantics.
(b) The global constraints need to be consistent with the integrated data from sources.
(c) There is a need to find the non-equivalent constraints between the source derived constraints at the global site and the constraints defined on the global site.
(d) The correspondences between relational constraints and XML constraints need to be investigated as there are many proposals on XML constraints [39] while relational constraints are well established.

5. Conclusions

We reviewed the constraints in data transformation and integration for different data models for information systems. We then showed what should be the architecture for data transformation according to the review. We also showed the data integration framework that used the transformations of schema, data and constraints.

References

[1] Lenzerini M.: Data Integration: A Theoretical Perspective. In: ACM PODS, pp. 233-246(2002)

[2] Poggi A.,Abiteboul S.: XML Data Integration with Identification. In: DBPL,pp.106-121(2005)

[3] Cali A., Calvanese D., Giacomo G. D., Lenzerini M.: Data Integration under Integrity Constraints. In: CAISE, pp.262-279(2002)

[4] Amer-Yahia S., Du F., Freire J.: A comprehensive solution to the XML-to-relational mapping problem. In: WIDM,pp.31-38(2004)

[5] Shahriar Md. S., Liu J.:Preserving Functional Dependency in XML Data Transformation. In: ADBIS, LNCS 5207, pp. 262-278(2008)

[6] Shahriar Md.S., Liu J.: Towards the Preservation of Keys in XML Data Transformation for Integration. In: COMAD, pp. 116-126(2008)

[7] Tim Bray, Jean Paoli, and C. M. Sperberg-McQueen, Extensible Markup Language (XML) 1.0., World Wide Web Consortium (W3C), Feb 1998. `http://www.w3.org/TR/REC-xml`.

[8] Henry S. Thompson, David Beech, Murray Maloney, and Noah Mendelsohn, XML Schema Part 1:Structures, W3C Working Draft, April 2000. `http://www.w3.org/TR/xmlschema-1/`.

[9] Liu J., Park H., Vincent M., Liu C.: A Formalism of XML Restructuring Operations. In: ASWC, LNCS 4185, pp. 126-132(2006)

[10] Ramesh V., Ram S.: Integrity Constraint Integration in Heterogeneous Databases: An Enhanaced Methodology for Schema Integration. In: Information Systems, Voll.22, No.8, pp.423-446(1999)

[11] Zamboulis L., Poulovassilis A.: Using Automed for XML Data Transformation and Integration. In: DIWeb, pp.58-69(2004)

[12] Zamboulis L.: XML Data Integration by Graph Restructuring. In: BNCOD, pp.57-71(2004)

[13] Su H., Kuno H., Rudensteiner E. A.: Automating the Transformation of XML Documents. In: WIDM, pp.68-75(2001)

[14] Erwig M.: Toward the Automatic Derivation of XML Transformations. In: ER, pp.342-354(2003)

[15] Li C.: Describing and utilizing Constraints to Answer Queries in Data Integration Systems. In: IIWeb(2003)

[16] Christiansen H., Martinenghi D.: Simplification of Integrity Constraints for Data Integration. In: FoIKs, LNCS 2942,pp.31-48(2004)

[17] Turker C., Saake G.: Consistent Handling of Integrity Constraints and Extensional Assertions for Schema Integration. in: ADBIS, LNCS 1691, pp.31-45(1999)

[18] Cali A., Calvanese D., Giacomo G. D., Lenzerini M.: On the Expressive Power of Data Integration Systems. In: ER, LNCS 2503,pp.338-350(2002)

[19] Cali A., Calvanese D., Giacomo G. D., Lenzerini M.: On the Role of Integrity Constraints in Data Integration. In: Bulletin of the IEEE Computer Society Technical Committee on Data Engineering(2002)

[20] Fuxman A., Miller R. J.: Towards Inconsistency Management in Data Integration Systems. In: IIWeb(2003)

[21] Popa L., Velegrakis Y., Miller R. J., Hernandez M. A., Fagin R.: Translating the web data. In: VLDB, pp. 598-609(2002)

[22] Davidson S., Fan W., Hara C., Qin J.: Propagating XML Constraints to Relations. In: ICDE, pp. 543-554(2003)

[23] Lee D., Chu W. W.: Constraint Preserving Transformation from XML Document Type Definition to Relational Schema. In: ER,LNCS 1920, pp. 323-338(2000)

[24] Liu Y., Zhong H., Wang Y., XML Constraints Preservation in Relational Schema. In: CEC-East(2004)

[25] Wang Q., Wu H., Xiao J.,Zhou A.: Deriving Relation Keys from XML Keys. In: ADC(2003)

[26] Liu Y., Zhong H., Wang Y.: Capturing XML Constraints with Relational Schema. CIT(2004)

[27] Liu C., Vincent M., Liu J.: Constraint Preserving Transformation from Relational schema to XML Schema. In: World Wide Web: Internet and Web Information Systems, 9, 93-110(2006)

[28] Deutsch A., Tannen V.: MARS: A System for Publishing XML from Mixed and Redundant Storage. In: VLDB(2003)

[29] Bertino E., Ferrari E.: XML and Data Integration. In: IEEE internet computing, pp. 75-76(2001)

[30] Benedikt M., Chan C.Y., Fan W., Freire J., Rastogi R.: Capturing both Types and Constraints in Data Integration. In: Sigmod(2003)

[31] Barbosa D., Freire J., Mendelzon A. O.: Information Preservation in XML-to-Relational Mappings. In: XSym, LNCS 3186,pp. 66-81(2004)

[32] Jiang H., Ho H., Popa L., Han W.: Mapping-Driven XML Transforamtion. In: WWW,pp. 1063-1072(2007)

[33] Fan W.: XML Constraints: Specification, Analysis, and Applications. In: DEXA, pp.805-809(2005)

[34] Fan W., Simeon J.: Integrity constraints for XML. In: PODS, pp.23-34(2000)

[35] McBrien P., Poulovassilis A.: A formalisation of semantic schema integration. In: Information Systems, 23 , pp. 307-334(1998)

[36] McBrien P., Poulovassilis A.: Schema Evolution in Heterogeneous Database Architectures,A Schema Transformation Approach. In: CAiSE, LNCS(2002)

[37] Poulovassilis A., Brien P. M.: A General Formal Framework for Schema Transformation. In: Data and Knowledge Engineering, 28(1998)

[38] McBrien P., Poulovassils A.: Data Integration by Bi-Directional Schema Transformation Rules. In: ICDE, pp. 227-238(2003)

[39] Hartmann S., Khler H., Link S., Trinh T., Wang J.: On the Notion of an XML Key. In: SDKB, pp. 103-112(2008)

Biography

Md. Sumon Shahriar: Sumon Shahriar is currently PhD researcher in Data and Web Engineering Lab, School of Com puter and Information Science, University of South Australia. He achieved his Bachelor of Science (Honours) and Master of Science (Research) degrees both with first class in Computer Science and Engineering from University of Dhaka, Bangladesh. His research interests include XML database, Data Integration, Data Quality and Data Mining.

Dr. Jixue Liu: Jixue Liu got his bachelor's degree in engineering from Xian University of Architecture and Technology in 1982, his Masters degree (by research) in engineering from Beijing University of Science and Technology in 1987, and his PhD in computer science from the University of South Australia in 2001. His research interests include view maintenance in data warehouses, XML integrity constraints and design, XML and relational data, constraints, and query translation, XML data integration and transformation, XML integrity constraints transformation and transition, and data privacy.