1. Introduction

The rapid growth of the World Wide Web and its technologies has resulted in enormous amounts of data being used over the Internet by Web Services and Web-based applications. The increase in semistructured data usage is not limited to Web applications but expands into various other applications such as digital libraries, biological databases, and multimedia data management systems. This expansion of semistructured data usage creates the need for effective and efficient utilization of semistructured data [15].

With such a rapid increase in its usage semistructured data needs to be stored, manipulated, and queried to be utilized properly by various applications and tools. For these purposes, many researchers have proposed to design and develop adequate database systems for semistructured data. As a result, several database systems have already been developed for eXtensible Markup Language (XML) [6], which is a common representation for semistructured data, while traditional database companies, such as Oracle, have provided XML support for their existing database systems.

As with widely used database systems, various operations that transform the schema have been adopted by the database systems developed for semistructured data to provide effective and efficient data storage and utilization. These schema transforming operations are often performed using algorithms developed specifically for semistructured data storage. The schema transforming operation guided by the algorithms must perform correctly to ensure the consistency of the data and ensure no information is lost. Although the algorithms claim to maintain the lossless and dependency preserving properties, the database systems developed for semistructured data lack verification support to prove the correctness of the transformations.

In widely adopted database systems, one of the features that is used to prove the correctness of the operations and algorithms is the mathematical foundation. For example, in relational database systems, a mathematical foundation has been extraordinarily useful in the definition of normalization, to prove that lossless and dependency proving algorithms can be defined. Also a mathematical foundation has been defined to capture object oriented concepts, and used to reason about the correctness of query results in object oriented database systems. Such verification support for operations and the algorithms of database systems ensures the correctness of data manipulations, making the mathematical foundation essential.

However, current developments of database systems that store semistructured data lack a mathematical foundation. When there is no general formal way of distinguishing between correct and incorrect transformations, incorrect data transformation can be introduced resulting in unreliable or even corrupt data. Without establishing a well defined mathematical foundation, many limitations will be imposed on the functionality of the database systems for semistructured data making it not as effective and reliable as it should be.

Therefore this research proposes to establish a well defined mathematical foundation for semistructured data in order to address this problem. The derived mathematical foundation will verify whether operations and algorithms that transform the schema of semistructured data maintain the lossless and dependency preserving properties.

2. Proposed Research

The main objective of this research is to establish a well defined mathematical foundation for semistructured data to verify the operations and algorithms that transform schemas of semistructured data. The mathematical foundation for semistructured data should contain formally specified semantics of a data modeling language with various database operations and algorithms represented accordingly. Also the mathematical foundation must be supported with adequate formal verification tools for verification of various database operations and algorithms. Hence, to define a mathematical foundation for semistructured data, a standard representation of schemas and the schema transformation
operators for semistructured data must be formally specified and verified using adequate data modeling languages and formal languages.

There has been other research that provides a formal semantics for semistructured data. For example, the formalization of DTD (Document Type Definition) and XML declarative description documents using expressive description logic has been presented by Calvanese et al. [3]. Anutariya et al. presented the same formalization using a theoretical framework developed using declarative description theory [1]. Also spatial tree logics have been used to formalize semistructured data by Conforti and Ghelli [4]. More recently, hybrid multimodal logic was used to formalize semistructured data by Bidoit et al. [2]. While these works have helped us develop a better understanding of the semantics of semistructured data, none of them have applied adequate and automated verification. Furthermore, none of these researchers have considered providing specification and verification for operations and algorithms that transform semistructured data schema.

As a result of examining related work and intensive background research, Object Relationship Attribute model for semistructured data (ORA-SS) data modeling language [5, 9] will be adopted as a data modeling language. The ORA-SS data modeling language is used because it not only captures the constraints that are represented in textual languages such as XML Schema [12] but also it is a diagrammatic notation which can be used for conceptual modeling.

With ORA-SS, we also applied a similar approach to formalize semistructured data using Z/EVEs [8] and Alloy [14]. But the approach using Z/EVEs had problems with complicated and time consuming verification and the approach using Alloy had a scalability problem. Considering these problems, the research will use Prototype Verification System (PVS) [10] and its theorem prover as the formal specification language and verification tool. Also PVS has proven its effectiveness by providing precise formal definitions and powerful automated verification support in various other research projects [11, 7, 13].

With ORA-SS data modeling language and PVS we will conduct the following tasks to complete the mathematical foundation for semistructured data.

- Specifying formal semantics of ORA-SS data model using PVS formal specification language with automatic verification support
- Verifying the defined ORA-SS formal semantics
- Specifying basic transformation operators according to the specification of ORA-SS formal semantics
- Representing existing database operations and corresponding algorithms for semistructured data such as normalization and view
- Verifying the correctness of the operations and algorithms (whether they maintain lossless and dependency preserving property)

When the formal definitions and verification described above are successfully completed, the research will result in the establishment of a well defined mathematical foundation for semistructured data. Additionally, the research can compare different algorithms derived for each database concept based on its performance to find the best algorithms using the verified representations of the database concepts.

This part of the research extends the formal mathematical foundation even further by enabling its practical applications on utilization of semistructured data in various applications and in its database systems. Also by representing these algorithms and verifying their correctness, the research will demonstrate the correctness and applicability of the mathematical foundation.

At the completion of all these tasks, the research will have defined a mathematical foundation and demonstrated its applications. The defined mathematical foundation of the research, that consists of verified formal specification of ORA-SS data model semantics, incorporated schema transformation operators and the verified representation of the best algorithms for each database operations, will be powerful enough to support effective and efficient use of semistructured data in various applications as well as its database systems. In addition, the research will also help the ORA-SS data model to evolve and provide a possibility for real Web-based applications to be developed from the ORA-SS data model.

3. Conclusions

This proposed research will establish a mathematical foundation for semistructured data to verify the schema and data transforming operations and algorithms for semistructured data. The advantages of having such a mathematical foundation includes providing formal semantics for semistructured data design, enhancing the discovery of inconsistencies in the data, providing verification for correctness of database operations such as normalization and view definitions, and maintaining lossless and dependency preserving properties of algorithms for database systems. Also the generic nature of the defined mathematical foundation allows it to be applied to any applications or database systems that use semistructured data.

Currently, the formal semantics of ORA-SS data model language has been specified and verified using PVS and its verification support. Using this formally specified and verified ORA-SS semantics, basic transformation operators will be defined and verified. Furthermore, based on the defined
semantics of the ORA-SS data model and basic transformation operators various database concepts and its algorithms will be defined and verified completing the mathematical foundation for semistructured data. Then the defined mathematical foundation will be evaluated through several case studies of conducting verification for some essential database operations such as normalization and view definitions and their algorithms.

According to the related work and preliminary research, the proposed research is evaluated to be unique and very valuable. The outcome of the research will surely eliminate the current shortcomings of database systems for semistructured data and provide tremendous benefits to the database systems for semistructured data.

Therefore it is believed that the contribution of this research is essential for the dramatically expanding semistructured data to be as powerful and versatile as existing structured data providing a solid basis for less structured information to be used in various applications.

References