Dependencies in Complex-value Databases

Sebastian Link

A dissertation presented in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Information Systems at Massey University

Supervisor:
Prof. Dr. Klaus-Dieter Schewe

Co-Supervisor:
Prof. Dr. Mike Hendy

Internal Examiner:
Associate Prof. Dr. Sven Hartmann

New Zealand Examiner:
Prof. Dr. Robert Goldblatt

Overseas Examiner:
Prof. Dr. Joachim Biskup

Date of Examination:
15.12.2004

Abstract

The relational data model has been the dominant model in database design for more than three decades. It considers data to be stored in matrices where rows correspond to individuals, columns correspond to attributes, and every cell constains a single atomic value. However, today's database technology trends, e.g. spatial, genetic or web-based data, require extended data models. Within the last decade new, complex-value data models such as the higher-order entity-relationship model, object-oriented data models, semi-structured data models, and XML have evolved which allow cells to contain lists, sets, multisets, trees, matrices or even more complex type constructors, references to other cells (which lead to infinite structures), and null values (indicating missing, unknown or vague data).

Matrices as such allow the storage of inconsistent data, invalid in the semantic sense. As this is not acceptable, additional requirements called dependencies have to be formulated when designing a database. The correct specification and use of dependencies needs a sound mathematical basis. For the relational data model more than 90 different classes of dependencies have been defined and studied intensively. The major problems in dependency theory are the axiomatisability of classes of dependencies, determination of the closure of a chosen set of dependencies (as certain dependencies can be implied by others) and the characterisation of semantically desirable properties for well-designed databases (such as absence of redundancies or abnormal update behavior) by syntactic properties on closed sets of dependencies.

With few exceptions research has only dealt with dependencies for the relational data model. Only recently, the emergence of XML as the standard format for web-based data and the rapidly increasing usage of persistent XML databases revealed the lack of a sound mathematical basis for complex-value data models. If they are expected to serve as first class data models they require a theoretical investigation of issues like integrity, consistency, data independence, recovery, redundancy, access rights, views and integration. The goal of this thesis is to develop a dependency theory for complex-value databases that is independent from any individual data model. Therefore, an abstract algebraic approach is taken that can be adapted to the presence of different combinations of type constructors such as records, lists, sets and multisets. Data models are classified according to the data types they support. In this framework the major objective is to initiate research on the following problems

- investigate the axiomatisation of important dependency classes, relevant to complex-value data models, by sound and complete sets of inference rules that permit the determination of all dependencies implied by some chosen set of dependencies.
- characterise semantically desirable properties by normal forms for complex-value data models and investigate whether these normal forms can always be achieved without violating other desirable properties.
- develop efficient algorithms for determining the closure of a chosen set of dependencies and for restructuring databases such that normal forms are satisfied and no information is lost.

In a single thesis it is impossible to consider all classes of relational dependencies in all different combinations of type constructors. Therefore the focus is put on extending two popular classes of relational dependencies: functional and multi-valued dependencies. The axiomatisation and implication of functional dependencies is investigated for all combinations of record, list, set and multiset type. Furthermore, a normal form with respect to functional dependencies in the presence of records and lists is proposed and semantically justified. It is also shown how to obtain databases which are in this normal form. Finally, axiomatisation and implication for the class of multi-valued dependencies and the class of functional and multi-valued dependencies are studied in the context of records and lists. The work of this thesis may lead to a unified dependency theory for complex-value data models.

Acknowledgement

I would like to thank *Klaus-Dieter Schewe* for giving me the opportunity to pursue an academic career under his supervision. Klaus-Dieter introduced me to theoretical computer science, and in particular to the topic of databases. His enthusiastic attitude towards research has motivated me since the beginning of my studies. Klaus-Dieter suggested to look at dependencies in the Higher-Order Entity-Relationship model, and it was later on that the investigations resulted in a more general treatment. The idea of classifying data models according to the data types they support is similar to the idea from [243] where object-oriented data models are classified according to different type systems. I also like to thank Klaus-Dieter for showing me how to work scientifically and for supporting decisions to reduce my teaching workload at critical times.

My thank also goes to *Mike Hendy* who kindly agreed to act as co-supervisor of this thesis.

Special thanks go to *Sven Hartmann* for having numerous fruitful discussions on the subject of this work. Sven made a lot of suggestions for improving the outcome, in particular on the mathematical rigorousness of various notions and proof arguments. Sven has always been prepared to talk about any issues, even in times he was very busy. I have learned a great deal from him.

I would further like to acknowledge *Bernhard Thalheim* and *Joachim Biskup* whose great experiences in the field of databases I have benefited from a lot.

Finally, I would like to thank my parents *Karla* and *Hans-Jürgen Link* who have always been there for me and supported me in every way possible.

Moreover, I am very thankful to my partner *Toni Floyd* who can always solve all my problems at once just by being who she is.

I dedicate this thesis to my parents,
Karla and *Hans-Jürgen Link*.

Table of Contents

List of Figures