Kurzfassung der Habilitationsschrift

Gesicherte und mögliche Antworten auf Anfragen an relationale Datenbanken mit partiellen Relationen

Hans–Joachim Klein
Institut für Informatik und Praktische Mathematik
Christian-Albrechts-Universität Kiel

In seinem Buch 'The Theory of Relational Databases' schrieb David Maier 1983 in der Einleitung zu einem Kapitel, das sich mit unvollständiger Information und Datenbanksemantik beschäftigt:
The reader is warned that the topics in this chapter are matters of personal taste. Whether the definitions and approaches offered seem right depends on individual intuition.

Trotz zahlreicher Arbeiten, die sich im Umfeld des relationalen Datenmodells mit der Darstellung und Auswertung unvollständiger Information auseinandergesetzt haben, müssen die mit den notwendigen Abweichungen vom klassischen Modell verknüpften Probleme auch heute noch als nicht befriedigend gelöst angesehen werden. Die unzureichend entwickelten theoretischen Grundlagen führen zu Schwierigkeiten bei der korrekten Handhabung unvollständiger Information in der Praxis, wie sie etwa beim Einsatz der Sprache SQL zu beobachten sind.

Die vorliegende Arbeit versucht, einen Beitrag zum wohldefinierten Umgang mit unvollständiger Information in relationalen Datenbanken zu leisten. Sie konzentriert sich auf zwei Formen unvollständiger Information, die Ursachen für das Fehlen gewöhnlicher Attributwerte in Relationen sein können und häufig in Anwendungen auftreten:
- die Existenz eines Wertes ist bekannt, aber der konkrete Wert ist unbekannt bzw.
- es liegt keinerlei Information zu einem Wert vor.

Die Untersuchungen werden im wesentlichen am klassischen relationalen Datenmodell mit flachen Relationen durchgeführt. Dabei stehen zwei unterschiedliche Erweiterungen gewöhnlicher Relationen im Mittelpunkt: ω-Relationen und V-Relationen. In ω-Relationen dient ein einziges ausgezeichnetes Symbol dazu, Vorkommen unbekannter Attributwerte anzuzeigen; in V-Relationen stehen Variable zur Verfügung, um jedes Vorkommen eines unbekannten Wertes eindeutig zu kennzeichnen. Neben diesen beiden Relationsformen sind im Zusammenhang mit der Diskussion der Nullwertproblematik der Sprache SQL auch SQL-Tabellen Gegenstand der Betrachtungen.

Nach Einführung einiger grundlegender Begriffe und Sprechweisen werden zunächst unterschiedliche Formen gesicherter Antworten für die erste der oben genannten Bedeutungen fehlender Werte vorgestellt. In gesicherten Antworten ist nur Information enthalten, die für die entsprechende Anfrage bei beliebiger Ersetzung aller Vorkommen unbekannter Attributwerte durch konkrete Werte gültig ist; dagegen heißt eine Antwort möglich, wenn sie bei mindestens einer solchen Ersetzung erhalten werden kann. Gesicherte Antworten eignen sich in unterschiedlicher Weise zur Beschreibung der im allgemeinen unendlichen Menge möglicher Antworten. Für die Darstellung von Antworten finden ω- und V-Relationen sowie Mischformen beider Relationsarten Verwendung. Die Definitionen basieren auf der Semantik von Anfragen in total definierten Zuständen. Sie sind unabhängig von einer konkreten Anfragesprache und nutzen Ordnungen, die sich auf partiell definierten Tupeln über deren Informationsgehalt ergeben. Durch Angabe geeigneter Interpretationsvorschriften für Antwortrelationen wird damit die Semantik von Antworten ohne Bezug auf konkrete Auswertungsmethoden eindeutig festgelegt.

Für die Relationenalgebra und den tupelorientierten Relationenkalkül untersuchen wir in den nachfolgenden Kapiteln, wie Operationsdefinitionen bzw. Vorschriften für die Interpretation von Kalkülausdrücken so gewählt werden können, daß eine effiziente Bestimmung gesicherter Antworten möglich ist. Dabei diskutieren wir zunächst entsprechende Vorschläge aus der Literatur. Wegen des mit unbekannten Werten verknüpften Tautologieproblems bei der Auswertung Boolescher Ausdrücke ist keine Vollständigkeit von Antworten zu erwarten, wenn Effizienz gewährleistet sein soll. Unter der Annahme, daß Vorkommen unbekannter Werte in ω-Relationen eine beliebige Anzahl von Werten repräsentieren, stellt sich sogar heraus, daß auch einfache Formen gesicherter Antworten nicht berechenbar sind.

Um die verschiedenen Formen gesicherter Antworten möglichst gut zu approximieren, werden für die Relationenalgebra jeweils zwei Varianten von Definitionen für die einzelnen Operationen eingeführt und für den Relationenkalkül mehrwertige Logiken verwendet. Die im Fall totaler Relationen geltende Äquivalenz von Relationenalgebra und Relationenkalkül kann auch für derartige Erweiterungen bewiesen werden. Darüberhinaus wird gezeigt, daß ein effizientes Verfahren, das im Zusammenhang mit der beweistheoretischen Sichtweise relationaler Datenbanken von R. Reiter vorgeschlagen wurde, zu den gleichen Ergebnissen führt wie ein in der Arbeit entwickeltes Verfahren für V-Relationen. Ein anderer Ansatz in der Literatur, der Vollständigkeitsprobleme durch Verwendung intuitionistischer Logik zu lösen versucht, erweist sich dagegen als nicht geeignet für die Bestimmung gesicherter Information in Antworten.

Obwohl gelegentlich als einfachere Variante bezeichnet zeigt sich die zweite der oben genannten Bedeutungen fehlender Werte, die auch den 'Nullwerten' von SQL zugrunde liegt, als die schwierigere Form, wenn es um die Festlegung gesicherter Information zu Anfragen geht. Einige Ansätze in der Literatur erweisen sich hier als nicht geeignet. Wir diskutieren die Probleme eingehend an Beispielen und machen einen Vorschlag zur Festlegung der Semantik von Datenbankzuständen, in denen für fehlende Werte nicht bekannt ist, ob sie definitiv nicht vorhanden sind oder nur keine konkreten Werte angegeben werden können.

Im letzten Kapitel wenden wir einen Teil der entwickelten Methodik auf die Sprache SQL an. Wir zeigen die Probleme auf, die durch den Übergang von einer dreiwertigen Logik zur Booleschen Logik bei geschachtelten Anfragen bezüglich gesicherter Information in Antworten sowie der Äquivalenz von Anfragen auftreten können. Für zwei Antwortformen, eine gesicherte und eine sogenannte potentielle, die sich beide auf die SQL-Semantik stützen, geben wir an, wie SQL-Anfragen zu transformieren sind, um stets Antworten der gewählten Form zu erhalten.