BestMasters

Mit „BestMasters“ zeichnet Springer die besten Masterarbeiten aus, die an renommierten Hochschulen in Deutschland, Österreich und der Schweiz entstanden sind. Die mit Höchstnote ausgezeichneten Arbeiten wurden durch Gutachter zur Veröffentlichung empfohlen und behandeln aktuelle Themen aus unterschiedlichen Fachgebieten der Naturwissenschaften, Psychologie, Technik und Wirtschaftswissenschaften.
Die Reihe wendet sich an Praktiker und Wissenschaftler gleichermaßen und soll insbesondere auch Nachwuchswissenschaftlern Orientierung geben.

Jürgen Hölsch

Optimierung von Nested Queries unter Verwendung der NF²-Algebra

Mit einem Geleitwort von
Prof. Dr. Michael Grossniklaus
und Prof. Dr. Marc H. Scholl

Springer Vieweg

Jürgen Hölsch
Konstanz, Deutschland

BestMasters
ISBN 978-3-658-12609-4 ISBN 978-3-658-12610-0 (eBook)
DOI 10.1007/978-3-658-12610-0

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Springer Vieweg

Gedruckt auf säurefreiem und chlorfrei gebleichtem Papier

Springer Fachmedien Wiesbaden ist Teil der Fachverlagsgruppe Springer Science+Business Media (www.springer.com)

Institutsprofil

Jürgen Hölschs Arbeit ist am Fachbereich für Informatik und Informationswissenschaft der Universität Konstanz entstanden, der derzeit dreizehn Professuren umfasst. Die Arbeitsgruppen an unserem Fachbereich beschäftigen sich insbesondere mit Methoden und Systemen zur Visualisierung, Analyse, Exploration und Verarbeitung von großen Informationsmengen. Dieser gemeinsame Forschungsschwerpunkt zeichnet die Forschung des Fachbereichs aus und ist in dieser Form einzigartig.

Betreut wurde die Arbeit von Jürgen Hölsch in der Arbeitsgruppe für Datenbanken und Informationssysteme (DBIS), die mit Marc H. Scholl und Michael Grossniklaus aus zwei Professuren sowie derzeit zwei Postdocs und fünf Doktoranden besteht. In der Vergangenheit lagen Forschungsschwerpunkte der Arbeitsgruppe u.a. auf den Gebieten XML-basierte Datenbanksysteme, Data Warehousing, Textdatenbanken und Integration von Compiler- und Datenbanktechniken. Dabei standen immer Nutzungsaspekte (Sprachen, Modelle, Schnittstellen) und Systemaspekte (Architektur, Performance, Optimierung) gleichermaßen im Fokus. Eine Vorgängerversion der in dieser Arbeit verwendeten NF^2-Relationenalgebra geht auf eigene Vorarbeiten aus den 1980er Jahren zurück.

Aktuell widmen sich unsere Arbeiten der Anfrageverarbeitung und -optimierung in verschiedenen Anwendungsgebieten. Ein solches Anwendungsgebiet ist die Verwaltung und Verarbeitung von XML-Daten. Im Rahmen der open-source XML-Datenbank BaseX, deren Entwicklung von einem Spin-off der DBIS-Gruppe vorangetrieben wird, werden neue Optimierungstechniken für die XML-Anfragesprache XQuery 3.0 untersucht. Da es sich bei XQuery 3.0 um eine komplette funktionale Programmiersprache handelt, ist es dazu notwendig Techniken von optimierenden Compilern mit denen von Anfrageoptimierern zu verbinden. Die daraus gewonnenen Erkenntnisse sollen dann in einem nächsten Schritt auf andere Anfragesprachen angewendet werden, die ebenfalls Konzepte aus Anfrage- und Programmiersprachen vermischen, wie zum Beispiel PL/SQL.

Ein anderes Anwendungsgebiet, dessen wir uns angenommen haben, ist die Verarbeitung von Textdatenströmen. Insbesondere werden hier Techniken untersucht, wie aus Social Media Datenströmen, wie sie beispielsweise von Twitter erzeugt werden, Ereignisse erkannt werden können. In diesem Zusammenhang spielt sowohl die Laufzeit der untersuchten Verfahren als auch die Qualität der berechneten Resultate eine große Rolle. Um in der Zukunft adaptive Ereigniserkennungsverfahren zu entwickeln, die sich dem veränderlichen Volumen eines Datenstroms anpassen können, untersuchen wir den genauen Trade-off zwischen Laufzeit und Resultatqualität von bekannten Ereigniserkennungstechniken.

Die Verarbeitung von Graph-Daten ist ein weiteres Anwendungsgebiet, in dem wir gegenwärtig neue Möglichkeiten der Anfrageoptimierung untersuchen. Insbesondere wollen wir für Sprachen, die in die Klasse $CRPQ^{agg}$ fallen, eine Algebra definieren. Im Rahmen dieser Algebra sollen dann Äquivalenzen gefunden werden, mithilfe derer ein transformierender kostenbasierter Optimierer im traditionellen Sinn entwickelt werden kann.

Die Arbeit von Jürgen Hölsch leistet einen direkten Beitrag zum ersten der beschriebenen Projekte, verkörpert aber gleichzeitig auch den Ansatz, den wir in den anderen beiden Projekten verfolgen, nämlich bekannte Ansätze – mit entsprechenden Erweiterungen – in neue Anwendungs- und/oder Systemkontexte zu integrieren.

Geleitwort

Die Arbeit von Jürgen Hölsch leistet einen signifikanten Beitrag zum Stand des Wissens im Bereich der Optimierung von SQL-Anfragen, einem wichtigen Teilgebiet der Datenbankforschung. Insbesondere nimmt sich diese Arbeit der Fragestellung an, wie verschachtelte SQL-Anfragen (*nested queries*) besser optimiert werden können. Dabei handelt es sich um ein sehr praxisrelevantes und aktuelles Thema. Einerseits sind verschachtelte Anfragen einer der wenigen Mechanismen, die einem SQL-Programmierer zur Verfügung stehen, um Anfragen modular und schrittweise aufzubauen. Aus diesem Grund wenden beispielsweise ungeübte SQL-Programmierer dieses Konzept auch dann an, wenn es zur Spezifikation der Anfrage eigentlich gar nicht nötig wäre. Andererseits sind die Optimierungstechniken, die heutige Datenbanksysteme auf verschachtelte Anfragen anwenden, eher simpel und lassen daher viele Optimierungsmöglichkeiten aus. Im schlimmsten Fall gelingt es diesen Systemen nicht, die (unnötige) Verschachtelung zu entfernen, was in der Regel zu inakzeptablen Ausführungszeiten führt. Speziell dieser zweite Punkt wird in dieser Arbeit in einer kleinen Studie gezeigt, indem die Optimierung und Ausführungszeit von eigens entwickelten verschachtelten Anfragen in vier großen Datenbanksystemen analysiert wird.

Zur besseren Optimierung von verschachtelten Anfragen wird ein Ansatz vorgeschlagen, der auf der sogenannten Non-First Normal Form (NF^2) Algebra beruht. Die NF^2-Algebra wurde in den 1980er-Jahren ursprünglich dazu entwickelt, Operationen in erweiterten relationalen Datenbanksystemen zu beschreiben, in denen verschachtelte Tabellen vorkommen können. Die Arbeit von Jürgen Hölsch zeigt, dass die NF^2-Algebra neben diesem ursprünglichen Anwendungsgebiet auch verwendet werden kann, um die Verschachtelungsmöglichkeiten darzustellen, die mittlerweile in SQL möglich sind. Aufbauend auf dieser Darstellung definiert die Arbeit in einem zweiten Schritt algebraische Äquivalenzen, die sowohl bekannte wie auch neue Optimierungstechniken für verschachtelte Anfragen formalisieren. Im Falle von neuen Techniken wird außerdem die Korrektheit der entsprechenden Äquivalenzen aufbauend auf den Grundlagen der NF^2-Algebra bewiesen. Der dritte Teil der Arbeit beschreibt die Implementation eines Anfrageoptimierers, der auf dem Cascades-Framework basiert und anstelle der traditionellen relationalen Algebra die NF^2-Algebra und ihre Äquivalenzen verwendet, um Anfragen zu transformieren. Da die NF^2-Algebra auf der relationalen Algebra aufbaut, können mit diesem Ansatz sowohl herkömmliche als auch verschachtelte Anfragen einheitlich optimiert werden. Dieser neuartige Anfrageoptimierer wird im letzten Teil der Arbeit qualitativ und quantitativ evaluiert, indem die von ihm umgeschriebenen Anfragen in drei kommerziellen und einem open-source Datenbanksystem ausgeführt und ihre Laufzeiten mit derjenigen der nicht voroptimierten Variante verglichen werden. Auf diesem Weg wird gezeigt, dass im Bereich von verschachtelten Anfragen noch großes Verbesserungspotential in den untersuchten Systemen besteht.

Michael Grossniklaus
Marc H. Scholl
Konstanz, Oktober 2015

Inhaltsverzeichnis