



Dissertation

zur Erlangung des akademischen Grades

Doktoringenieur (Dr.-Ing.)

QBäume – Effizientes Retrieval von Graphen mit Hilfe von
Strukturinvarianten

vorgelegt

der Fakultät für Informatik

der Technischen Universität Chemnitz

von: Ralf Neubert
geboren am: 17. August 1974
in: Karl-Marx-Stadt

Chemnitz, den 16. Juni 2007

Gutachter: Prof. Dr. rer. nat. Andreas Henrich
Prof. Dr. rer. nat. Andreas Goerdt
Betreuer: Prof. Dr. rer. nat. Wolfgang Benn

Neubert, Ralf

QBäume – Effizientes Retrieval von Graphen mit Hilfe von Strukturinvarianten

Dissertation, Fakultät für Informatik

Technische Universität Chemnitz, März 2008

verfaßt auf Grundlage der bis Juli 1998 geltenden Rechtschreibregeln

Berichte aus der Informatik

Ralf Neubert

**QBäume – Effizientes Retrieval von Graphen
mit Hilfe von Strukturinvarianten**

D 93 (Diss. TU Chemnitz)

Shaker Verlag
Aachen 2008

Bibliografische Information der Deutschen Nationalbibliothek

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

Zugl.: Chemnitz, Techn. Univ., Diss., 2007

Copyright Shaker Verlag 2008

Alle Rechte, auch das des auszugsweisen Nachdruckes, der auszugsweisen oder vollständigen Wiedergabe, der Speicherung in Datenverarbeitungsanlagen und der Übersetzung, vorbehalten.

Printed in Germany.

ISBN 978-3-8322-7166-4

ISSN 0945-0807

Shaker Verlag GmbH • Postfach 101818 • 52018 Aachen

Telefon: 02407 / 95 96 - 0 • Telefax: 02407 / 95 96 - 9

Internet: www.shaker.de • E-Mail: info@shaker.de

Danksagung

Ich möchte mich bei all denen bedanken, die auf verschiedenste Weise zum Gelingen meiner Dissertation beigetragen haben.

Großen Dank schulde ich Prof. Dr. Wolfgang Benn, ohne die mir von ihm überlassene wissenschaftliche Freiheit und die eingeräumten Entwicklungsmöglichkeiten hätte ich diese Dissertation nicht anfertigen können. Ähnliches gilt für meine Kollegen an der Professur Datenverwaltungssysteme der Technischen Universität Chemnitz, deren selbstverständliche Kooperations- und unendliche Hilfsbereitschaft, erst das notwendige Klima schuf, um meine wissenschaftlichen Interessen über all die Jahre hinweg verfolgen zu können. Es war ein Privileg und eine Freude in diesem Team arbeiten zu dürfen.

Mein besonderer Dank für den intensiven Gedankenaustausch und die anhaltende Ermutigung meine Ideen in einer Dissertation niederzulegen gilt Dr.-Ing. Otmar Görlitz. Den Diskussionen mit ihm verdanke ich viele Anregungen und Einsichten, den Ansporn meine Ideen praktisch und theoretisch zu beweisen sowie die Überwindung der Selbstzweifel, die einen beim wissenschaftlichen Schreiben immer wieder befallen. Für die technische Unterstützung der Entwicklungsarbeiten möchte ich mich bei meinen ehemaligen studentischen Mitarbeitern Christian Zielke und Torsten Zesch bedanken. Außerdem bedanke ich mich an dieser Stelle bei Herrn RA Michael Koitsch, der mit seinem juristischen Geschick dafür gesorgt hat, daß die Universitätsverwaltung mich nicht dauerhaft in meiner Arbeit behindern konnte. Ebenso danke ich ausdrücklich Frau Ute Fischer, die mich mit scharfen und konzentriertem Blick durch den Irrgarten der deutschen Rechtschreibregeln gelotst hat.

Nicht zuletzt danke ich meiner Familie und meinen Freunden. Ohne euren Rückhalt und eure Aufmunterung hätte ich die lange Zeit des Schreibens dieser Dissertation nicht ertragen können.

Chemnitz, im Frühjahr 2008

Ralf Neubert

Inhaltsverzeichnis

Abbildungsverzeichnis	v
Tabellenverzeichnis	vii
1 Einleitung und Motivation	1
2 Graphen	3
2.1 Elementare Grundbegriffe	3
2.2 Darstellung von Graphen	4
2.3 Subgraphen	6
2.4 Klassifikation von Graphen	8
2.5 Morphismen zwischen zwei Graphen	10
2.5.1 Isomorphie- und Subgraphisomorphieproblem	11
2.5.2 Komplexitätsbetrachtungen	13
2.5.3 Einsetzbare Algorithmen zur Lösung von IP und SIP	15
2.6 Isomorphie und Subgraphisomorphie als Relation in einer Graphmenge	21
2.7 Invarianten von Graphstrukturen	24
2.7.1 Beispiele für Invarianten	28
3 Retrieval von Graphen	31
3.1 Anwendungsbereiche	31
3.2 Datenbankunterstützung der Anwendungsbereiche	37
3.3 Abgrenzung der betrachteten Retrievalprobleme	38
4 Indexierungsansätze für Graphmengen	41
4.1 Indexierte Zugriffsstrategie	41
4.2 Indexierungsansätze	44
4.2.1 Featurevektorindexe	46
4.2.2 Preprocessing-Ansatz	48
4.2.3 Path-Hashing-Ansatz	52
4.2.4 Der gIndex-Baum	57
4.2.5 Invariantenansatz	62
4.3 Bewertung der Indexierungsansätze	65
5 Topologische Invarianten von Relationenstrukturen	69
5.1 Q-Analyse	69
5.2 Alternative Anwendungen der Q-Analyse	77
5.3 Berechnung des Q-Vektors für einfache Graphen	78

5.4	Anwendbarkeit der Q -Vektorinvariante für Isomorphieanfragen	85
5.5	Anwendbarkeit der Q -Vektorinvariante für Subgraphanfragen	91
6	Redefinition des Q-Vektors	95
6.1	Erweiterung des Q -Vektors zum Q^+ -Vektor	95
6.2	Anwendbarkeit der Q^+ -Vektorinvariante für Isomorphie- und Subgraphanfragen	100
6.3	Erweiterung des Q^+ -Vektors zum Q^{bc} -Vektor	109
6.4	Definition und Berechnung des Q^{bc} -Vektors	111
6.5	Anwendbarkeit der Q^{bc} -Vektorinvariante für Isomorphie- und Subgraphanfragen	118
6.5.1	Möglichkeiten zur Reduktion von falschen Positiven bei Subgraphanfragen	123
6.5.2	Analyse der Abbildungen zwischen den Strukturcodeklassen	127
6.6	Ergänzung der Invariantenvektoren durch Attributinformationen	131
6.7	Zwischenfazit	133
7	Untersuchungen zur Differenzierungsfähigkeit der Invarianten	135
7.1	Zählen von Graphstrukturen	136
7.2	Empirischer Ansatz auf Basis von Graphkatalogen	142
7.3	Ergebnisse der empirischen Untersuchungen	144
8	QBäume	157
8.1	Implementierungsanforderungen und -entscheidungen	157
8.2	Repräsentation der Schlüssel in einem QBaum	160
8.3	Physische Seitenstruktur der Indexebene	161
8.4	Physische Seitenstruktur der Datenebene	164
8.5	Einfügen von Graphen in den QBaum	165
8.6	Isomorphieanfragen an einen QBaum	168
8.7	Subgraphisomorphieanfragen an einen QBaum	170
8.8	Löschen von Graphen aus dem QBaum	172
9	Experimentelle Untersuchungen an QBäumen	175
9.1	Leistungsgrenze der Q -Vektorinvariante	175
9.1.1	Aufbau des Experiments	175
9.1.2	Durchführung und Ergebnisse	176
9.1.3	Auswertung	178
9.2	Vergleichende Performanzuntersuchungen	179
9.2.1	Allgemeiner Aufbau der Experimente	180
9.2.2	Graph- und Anfragemengen für die Versuche	181
9.2.3	Testergebnisse für den initialen Indexaufbau	183
9.2.4	Testergebnisse für Subgraphisomorphieanfragen	184
9.2.5	Testergebnisse für Isomorphieanfragen	185
9.2.6	Auswertung der Ergebnisse	187
	Ergebnisse des initialen Indexaufbaus	187
	Performanz der Verarbeitung von Subgraphisomorphieanfragen	189
	Performanz der Verarbeitung von Isomorphieanfragen	190

9.2.7	Schlußfolgerungen aus den Ergebnissen	191
10	Zusammenfassung	193
10.1	Resümee	193
10.2	Möglichkeiten der Weiterentwicklung und zukünftige Forschungsthemen	195
A	Datenstruktur eines Eintrags in einer Proteindatenbank	197
B	Weitere Ergebnisse der Untersuchungen zur Differenzierungsfähigkeit von Invarianten	199
	Literaturverzeichnis	205