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

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

QBäume – Effizientes Retrieval von Graphen mit Hilfe von Strukturinvarianten

Ralf Neubert
Technische Universität Chemnitz

Zusammenfassung:

Viele Anwendungen nutzen zur Darstellung ihrer Anwendungsobjekte attributierte Graphen. Besonders häufig sind solche Anwendungen in Bereichen wie z. B. Computer Vision, Natural Language Processing, Expertensysteme sowie Chemieinformationssysteme und Biomolekulardatenbanken zu finden.

Aus der Retrieval- bzw. Datenbankperspektive erfordern derartige Anwendungen neben den attributwertbasierten Zugriffsmethoden vor allem Zugriffsoperationen, die beim Vergleich der Anfrage mit den Datenobjekten auch den Strukturaspekt berücksichtigen. Insbesondere sind die Suche nach identischen Graphen oder Graphen, die einen bestimmten anderen Graphen enthalten, zu unterstützen. In der theoretischen Informatik sind die dafür notwendigen Vergleichsoperationen unter dem Namen Isomorphieproblem und Subgraphisomorphieproblem bekannt. Weiterhin ist bekannt, daß Algorithmen zur Lösung dieser beiden Probleme eine sehr hohe Berechnungskomplexität aufweisen, so daß sich das sequentielle Suchen in großen Sammlungen von Graphen aus Laufzeitgründen verbietet. Zur Lösung des Retrieval-Problems in den eingangs genannten Anwendungsfeldern sind folglich geeignete Indexstrukturen notwendig, welche die Vergleiche von Graphen auf ein notwendiges Minimum reduzieren.

Daher ist die äußerst interessante und vielschichtige Frage nach einer effizienten Verwaltungs- und Retrieval-Struktur zur Beantwortung der beschriebenen Suchanfragen das Kernthema des vorliegenden Buches. Wie der Titel verrät, werden zur Lösung des Problems Strukturinvarianten bemüht. Präziser formuliert, wird vorgeschlagen, die konzeptionellen Ideen der Q-Analyse zur Untersuchung topologischer Invarianten von Relationenstrukturen für die Indexierung zu nutzen. Die Invarianten verleihen der entwickelten Indexstruktur ihren Namen *QBaum*. Auf Basis der bei der Q-Analyse gewonnenen Strukturinvarianten können Indexstrukturen aufgebaut werden, die derzeit bekannten Graphindexen bezüglich ihrer Retrievalperformanz und ihres Speicherplatzbedarfs mindestens gleichrangig, in vielen Bereichen sogar überlegen sind.

Der Nachweis dieser Behauptungen wird schrittweise vollzogen. Ausgehend von einer Gegenüberstellung der Techniken und Strategien für die Indexierung von Graphmengen wird die Auswahl des Invariantenansatzes begründet. Die zentralen Kapitel sind thematisch der Darstellung der bereits erwähnten Q-Analyse sowie der detaillierten Vorstellung der daraus gewonnenen Invariantenfunktionen und ihrer Eigenschaften gewidmet. Insbesondere wird gezeigt, daß Isomorphieanfragen an eine Graphmenge mit Hilfe des Q-Vektors vollständig und korrekt beantwortet werden können. Außerdem werden zwei neue Redefinitionen der Q-Vektorinvariante präsentiert, um auch Subgraphisomorphieanfragen vollständig und korrekt beantworten zu können.

Desweiteren wird die Differenzierungsfähigkeit der drei Invarianten untersucht. Dabei werden sie mit Invarianten verglichen, die derzeit für die Indexierung verwandt werden, um festzustellen, ob sie das Potential zur Entwicklung performanterer Invariantenindexe, als bisher bekannt ist, besitzen.

Abschließend wird die Implementierung des QBaum als neue Verwaltungsstruktur auf Basis der drei Invarianten vorgestellt und einer vergleichenden Performanzuntersuchung unterzogen, die sie anderen Indexen gegenüberstellt. Dazu werden verschiedene Experimente mit künstlich generierten Graphmengen und aus dem Umfeld von chemischen Informationssystemen stammender Graphdaten präsentiert und ausgewertet. Die Ergebnisse belegen, daß es dem QBaum gelingt, Graphen besser als andere derzeit bekannte Graphindexe zu organisieren und das Retrieval von Graphen bei Subgraph- und insbesondere Isomorphieanfragen deutlich zu beschleunigen.

Q-Trees – Efficient Retrieval of Graphs Using Structural Invariants

Ralf Neubert
Chemnitz University of Technology

Extended abstract: Graphs are a powerful concept for representing objects in many engineering and science applications. Some applications, e. g. chemistry and molecular biology applications or pattern recognition and computer vision applications, maintain and search large databases of graphs.

The comparison of two graph objects requires graph-matching operations, namely graph and subgraph isomorphism matching. Both matching operations inherently possess a high computational complexity. The retrieval of graphs from a set of graphs, does not only require the comparison of a query graph with another single graph, but the matching of a query graph to an entire database of graphs. If such a database is large, the sequential search approach -- matching each graph from the database with the query graph -- becomes prohibitively slow and infeasible. This means, that indexed access methods are needed to speed up the retrieval.

This thesis studies the problem of indexing a set of labelled undirected graphs and speeding up isomorphism and subgraph isomorphism queries to this set. For this purpose vector-valued functions extracting isomorphism invariant properties from a graph's structure and its labels are investigated. The considered functions are called invariants. The general approach pursued is to associate each graph in the database with its corresponding vector and to use them as keys for a robust index structure. The retrieval process consists of three basic steps. A query vector is extracted from the query graph. By comparing the query vector with the vectors in the index structure a subset of candidate graphs is retrieved from the database. Finally, only the candidate graphs are checked by the computationally expensive graph-matching operations to determine the result set.

In order to be efficient the described indexing approach needs invariants that can be quickly computed and effectively allow to distinguish between the various graphs. The last mentioned property of an invariant is the crucial factor for selecting small candidate subsets and reducing the graph-matching operations towards a necessary minimum. Therefore several invariants have been analyzed and compared to each other in this work. Invariants based on the *Q-analysis function* turned out as the most effective among the studied functions. The Q-analysis procedure is a concept developed in set theory for characterizing the topological structure of binary relationships. Based on this concept an index structure, called *Q-Tree*, has been developed, which supports isomorphism queries as well as subgraph isomorphism queries.

The implementation of the Q-Tree was built upon a Prefix-B*-Tree performing key compression. The index has been successfully tested on synthetically generated data sets and on a real-world graph database, containing molecule structures and medical annotations. The experiments presented in the thesis evaluate the retrieval properties of the Q-vector invariant and the performances of the Q-Tree for both query types. Furthermore the Q-Tree's performance is compared to another recently proposed index structure called GraphGrep. This access structure uses the concept of path hashing as general indexing approach. It is shown that the Q-Tree reduces the average time for answering a graph query by one to two orders of magnitude compared to path-hashing based indexing methods, while causing significantly less overhead with respect to insertion time and storage space.