Beitrag 8

Hybride Indexstrukturen

Carsten Kropf

Professur für Produktionswirtschaft und Informationstechnik
carsten.kropf@iisys.de

Abstract: Im Folgenden wird ein Promotionsprojekt zur Implementierung und Optimierung von hybriden Indexstrukturen beschrieben. Die erhöhte Suchperformance wird bei hybriden Indexstrukturen durch einen höheren Aufwand an Vorberechnungen bei Einfügeoperationen erreicht. Dadurch ergibt sich, im Gegensatz zu Ansätzen, welche mehrere Indexstrukturen miteinander verbinden oder getrennte Suchanfragen ausführen eine Effizienz der Reorganisation hybrider Indexstrukturen, die prohibitiv für den Einsatz in den meisten Anwendungen ist. Diese sollen innerhalb des Promotionsprojekts optimiert werden, um eine Einsatzfähigkeit in realistischen Szenarien gewährleisten zu können.

Einführung

In heutigen betrieblichen Informationssystemen, wie z.B. Dokumenten-Management-, Enterprise-Resource-Planning- oder Enterprise-Content-Management-Systemen, treten unterschiedliche Datentypen auf. Dabei kommen, vor allem bei der Verwendung von Dokumenten-Management-Systemen unstrukturierte oder nicht normalisierte Datentypen wie z.B. Texte vor, die mit Hilfe von angepassten Indexstrukturen (z.B. invertierter Index) in den darunter liegenden Datenbanksystemen durchsucht werden können. Enterprise-Resource-Planning Systeme hingegen verwalten generell relationale und normalisierte Datentypen, wie z.B. Preise oder Mengenangaben, die ebenfalls auf Basis von entsprechenden Indexstrukturen effizient durchsucht werden können. Bei der Verwendung von Enterprise-Content-Management-Systemen, welche die Integration verschiedener Systeme zum Ziel haben, müssen nun die Daten aus heterogenen Systemen integriert und durchsucht werden, um unternehmensweite Suchen ermöglichen zu können. Aktuell werden für den Einsatz von unternehmensweiten Suchen Konnektoren verwendet, die die unterschiedlichen Systeme unabhängig voneinander durchsuchen und am Ende die Schnittmenge, welche das finale Resultat der Suche darstellt, generieren. Sollen beispielsweise Informationen über Dokumente, die ausgewählte Schlüsselworte (als unstrukturiertes Kriterium) enthalten, innerhalb eines wohl definierten Zeitraums (als strukturiertes Kriterium) abgerufen werden, kann es vorkommen, dass beide unabhängigen Suchprädikate eine große Anzahl an Teilergebnissen generieren. Das hat zur Folge, dass zunächst große

Datenmengen aus den angeschlossenen Systemen geladen werden müssen, aber die endgültige Ergebnismenge nur relativ klein ist. Diese Teilmengen aus den jeweiligen Systemen können so groß sein, dass sie zunächst erneut auf der Festplatte zwischengespeichert und anschließend linear miteinander verglichen werden müssen. Dieser Prozess führt zu einem enormen zeitlichen Mehraufwand.
Hybride Indexstrukturen, die in den vergangenen Jahren entwickelt wurden (siehe u.a.: [GH+09], [GK10], [FHR08], [HH+07] oder [IBS08]), bilden Lösungsansätze für eine Art dieser kombinierten Suche. Diese Indexstrukturen unterstützen die kombinierte Indizierung verschiedenartiger Datentypen. Das bedeutet, dass hierbei innerhalb einer einzigen Indexstruktur nicht-relationale und relationale Daten effizient durchsucht werden können. Sie werden aktuell hauptsächlich im Bereich des *Geographic Information Retrieval*, also der kombinierten Suche nach Schlüsselworten und geographischen Regionen, eingesetzt, sind aber auf Grund ihrer flexiblen Struktur auch auf unternehmensbezogene Daten adaptierbar.
Die Suche innerhalb dieser Indexstrukturen stellt sich als effizient dar. Allerdings ergeben sich bei Reorganisationen, wie Einfüge-, Lösch- oder Datenmanipulationsoperationen, Probleme durch komplexe Abläufe und Zusammenhänge von Teilstrukturen untereinander. Dies hat zur Folge, dass die Reorganisationsoperationen entsprechend ineffizient sind.
Der Hauptfokus von betrieblichen Informationssystemen besteht darin, einem Anwender möglichst schnell die gewünschten Informationen präsentieren zu können. Bei einer ineffizienten Reorganisationsalgorithmik ergibt sich jedoch das Problem, dass die Ergebnisse einer Suche innerhalb dieser Systeme nicht schnell ausgeliefert werden können, da Reorganisationsalgorithmen die Indexstruktur blockieren und daher keine simultanen Suchanfragen ausgeführt werden können. Somit wird die Suchperformance durch blockierende Operationen innerhalb der Datenmanipulation signifikant beeinträchtigt. Dies führt dazu, dass hybride Indexstrukturen aktuell zwar theoretisch bezüglich des Suchaufwands eine Verbesserung darstellen, praktisch jedoch derzeit nicht einsetzbar sind.

Diese Problematik führt somit zu den Hauptfragestellungen, die sich bezüglich der Optimierbarkeit solcher Strukturen innerhalb eines Datenbanksystems, das als Grundlage für ein Informationssystem dient, beschäftigen.

Hauptfragestellungen

Ausgehend von den in der Einleitng geschilderten Problemen bezüglich der Effizienzbetrachtungen der aktuell vorliegenden Algorithmik ergeben sich nun die Hauptfragestellungen der Arbeit, die eng miteinander verbunden sind.

1. Welche (Teil-)Algorithmen der Indexstrukturen sorgen für die schlechte Performance beim Einfügen neuer, bzw. Manipulieren vorhandener Datensätze?
2. Kann die Effizienz der Algorithmen dahingehend gesteigert werden, dass eine hinreichend gute Performance während des Einfügens die Suchzeit nicht negativ beeinflusst?

Vorgehen und Lösungsansätze

Im Allgemeinen sollen in dieser Arbeit die ineffizienten (Teil-)Algorithmen, die für Performanceprobleme verantwortlich sind, ermittelt, analysiert und optimiert werden.
Bei diesem Vorgehen wird eine Forschungsmethodik, angelehnt an *Design Science* [HM+04] und *Regulative Cycle* [Wie09], verwendet. Der *Regulative Cycle* kommt insbesondere darin zum Tragen, dass eine Softwareoptimierung mit entsprechender Validierung in mehreren Iterationsschritten ausgeführt wird.
Im ersten Schritt wird zunächst ein Testsystem erstellt, welches auf Basis von wohldefinierten Testdaten die ineffizienten Algorithmen identifizieren kann. Dazu werden Datensätze benötigt, welche anschließend in einer Testumgebung verwendet werden können. Außerdem müssen zur Definition des Testsystems die zu messenden Merkmale der Software identifizert werden.
Als Datensätze für die auszuführenden Tests werden englischsprachige Wikipedia Artikel, Artikel der Nachrichtenagentur Reuters und synthetische Daten verwendet. Synthetische Datensätze bieten hierbei den Vorteil, bestimmte Auffälligkeiten, welche bei Messungen aufgetreten sind, gezielt überprüfen zu können. Außerdem kann noch auf die bereits analysierten Daten einer *Geographic Information Retrieval* Suchmaschine zurückgegriffen werden.
Die zu messenden Merkmale umfassen die für die jeweilige Algorithmen benötigte Zeit und weitere typische Merkmale für Indexstrukturen in Datenbanken, z.B. geladene Blöcke, Aufbau der Tupel, detaillierte Informationen über den aktuellen Zustand der Struktur etc. Mit Hilfe dieser Daten und eines entsprechenden Auswertungssystems werden nach einem Testlauf die ineffizienten (Teil-)Algorithmen innerhalb des Ablaufs der Reorganisationsalgorithmik der Indexstrukturen ermittelt. Basierend auf den generierten Vorschlägen werden anschließend Anpassungen an den entsprechenden Algorithmen vorgenommen, deren Effekte in der Folge wiederum mit Hil-

fe des Testsystems verifiziert werden. Die Daten der nach der Optimierungsphase durchgeführten Testphase gehen anschließend erneut in eine Auswertung ein, welche abermals als Basis für Anpassungen der Algorithmik genutzt wird. Aus diesem Vorgehen ergibt sich ein iterativer Prozess zur schrittweisen Optimierung der Reorganisationsalgorithmik für hybride Indexstrukturen. Hierbei ist zu beachten, dass die gemessenen Laufzeiten der Algorithmen nur als Hinweis für eine potentielle Optimierung stehen. Genauere Analysen über die Möglichkeiten zur Optimierung müssen in einem weiteren Schritt mit Hilfe angepasster Methoden, z.B. statischer Codeanalyse, durchgeführt werden.
Als Basis wird hierbei eine Implementierung genutzt, die auf bereits vorhandenen Methoden im Bereich der hybriden Indexstrukturen aufsetzt. Somit wird versucht, diese Implementierung dahingehend zu optimieren, dass „akzeptable" Laufzeiten der Reorganisationsalgorithmik unter verschiedenen Bedingungen mit unterschiedlichen Testdatensätzen erzielt werden können. Aktuell existente Indexstrukturen, wie B-Baum [BM72] oder R-Baum [Gut84], garantieren Reorganisationszeiten, die sich im Bereich von deutlich weniger als $1000ms$ bewegen. Dies führt zu einer uneingeschränkten Verwendungsmöglichkeit solcher Indexstrukturen innerhalb von Informationssystemen. Somit ist ein Ziel dieser Arbeit, einen ähnlich geringen zeitlichen Reorganisationsaufwand zu erreichen, um die praktische Einsetzbarkeit der hybriden Indexstrukturen innerhalb von betrieblichen Informationssystemen gewährleisten zu können. Um die Grenzen dieser Verwendungsmöglichkeit zu definieren, werden mehrere Textcorpora verwendet. Somit kann gezeigt werden, dass die entsprechende Performance auch unter verschiedenen Voraussetzungen erreicht werden kann. Durch die Aufteilung in mehrere Iterationen der Optimierung soll garantiert werden, dass die entsprechend umgesetzten Optimierungsansätze in der Tat einen effizienten Nutzen für die Reorganisationsalgorithmik haben. Außerdem kann darüber garantiert werden, dass, im Falle einer potentiellen Verschlechterung der Laufzeit, die entsprechenden Anpassungen problemlos rückgängig gemacht werden können.
Des Weiteren müssen alle gemessenen Werte aus entsprechenden Pre- oder Post-Tests mit Hilfe aller zur Verfügung stehenden Corpora validiert werden, um eine Optimierung auf einen bestimmten Corpus hin ausschließen zu können. Somit kann garantiert werden, dass die durchgeführten Adaptionen nicht nur einen bestimmten Sonderfall, sondern den allgemeinen Fall abdecken.

Literaturverzeichnis

[BM72] Bayer, R; McCreight, E M. Organization and Maintenance of Large Ordered Indexes. *Acta Informatica*, 1:173–189, 1972.

[FHR08] Felipe, I D; Hristidis, V; Rishe, N. Keyword Search on Spatial Databases. In *International Conference on Data Engineering*, S. 656-665, 2008.

[GH+09] Göbel, R, Henrich; A, Niemann, R; Blank, D. A hybrid index structure for geo-textual searches. In *Proceeding of the 18th ACM conference on Information and knowledge management*, CIKM '09, S. 1625-1628, ACM, New York, 2009.

[GK10] Göbel, R; Kropf, C. Towards Hybrid Index Structures for Multi-Media Search Criteria. In *Distributed Multimedia Systems*, S. 143-148, 2010.

[Gut84] Guttman, A. R-TREES. A DYNAMIC INDEX STRUCTURE FOR SPATIAL SEARCHING. In *SIGMOD '84: Proceedings of the 1984 ACM SIGMOD international conference on Management of data*, S. 47-57, ACM, New York, 1984.

[HH+07] Hariharan, R, Hore, B, Li, C; Mehrotra, S. Processing Spatial-Keyword (SK) Queries in Geographic Information Retrieval (GIR) Systems. In *SSDBM '07: Proceedings of the 19th International Conference on Scientific and Statistical Database Management*, S. 16, IEEE Computer Society, Washington, DC, USA, 2007.

[HM+04] Hevner, A R, March, S T, Park, J; Ram, S. Design science in information systems research. *Management Information Systems Quarterly.*, 28(1):75-105, 2004.

[IBS08] Ilyas, I F, Beskales, G; Soliman, M A. A survey of top-k query processing techniques in relational database systems. *ACM Computing Surveys*, 40(4):1–58, 2008.

[Wie09] Wieringa, R. Design science as nested problem solving. In *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology*, DESRIST '09, S. 8:1-8:12, ACM, New York, 2009.

Carsten Kropf, M.Eng., absolvierte zwischen 2004 und 2008 ein Studium zum Diplom Informatiker (FH) an der Hochschule Hof. Anschließend folgte ein Masterstudium im Studiengang Software Engineering for Industrial Applications. In der Masterthesis beschäftigte er sich mit dem Thema „A Hybrid Index Structure Supporting Multimedia Search Critiera". Diese Thesis führte weiter zum Promotionsprojekt, welches sich um die Optimierung der Reorganisationsalgorithmen dieser hybriden Indexstrukturen beschäftigt.

Dieser Beitrag ist erschienen in: Thorsten Claus und Niels Seidel (Hrsg.), *Werkstatt europäischen Denkens – 20 Jahre Internationales Hochschulinstitut Zittau*, TUDpress, Dresden, 2014. Online verfügbar: http://nbn-resolving.de/urn:nbn:de:bsz:14-qucosa-152297.