Exposé zur Diplomarbeit

Nutzung von Statistiken über Daten-Overlap zur Anfrageoptimierung in Peer Data Management Systemen

Véronique Tietz

Ein Peer Data Management System (PDMS) ist ein verteiltes Informationssystem, das aus einem Netzwerk von autonomen Quellen („Peers“) mit im Allgemeinen heterogenen Schemata besteht. In einem solchen System wird eine Anfrage von einem Peer bearbeitet, indem sie mithilfe von sogenannten Peer-Mappings an andere, benachbarte Peers weitergeleitet wird und die Ergebnisse geeignet zusammengeführt werden.

Je nach Verteilung der Daten in dem PDMS können diese Ergebnisse sowohl intensionale als auch extensionale Überlappungen bezüglich *Real-World*-Entitäten aufweisen. Während intensionaler Overlap die Integration von Daten überhaupt erst ermöglicht, führt extensionaler Overlap zu teilweise beträchtlichen Redundanzen sowohl im Datentransport als auch bei der Evaluierung der Anfrage. Hauptziel dieser Arbeit ist es daher, diese Redundanzen zu vermeiden und damit zu einer weiteren Optimierung der Anfragebearbeitung beizutragen, sodass letztendlich auch eine höhere Skalierbarkeit des PDMS erreicht werden kann.

Ausgehend von verschiedenen Topologien eines PDMS sollen hierzu im Rahmen der Arbeit zunächst mögliche Szenarien von Datenverteilungen über das gesamte PDMS und deren Auswirkungen auf den lokal beobachtbaren Daten-Overlap bzw. die verteilte Anfragebearbeitung analysiert und diskutiert werden.

Aufgrund der Autonomie der einzelnen Peers und des Fehlens einer zentralen Instanz mit „globalem Wissen“ über das PDMS kann ein einzelner Peer über die Datenverteilung bzw. den Overlap von Daten im Gesamtsystem keine direkte Aussage treffen. Darum ergibt sich die Herausforderung, die Anfrageantworten eines Mappings (also sogenanntes „query-feedback“) zu nutzen, um mithilfe von Statistiken auf die dahinterliegenden Daten zu *schließen*.

In dem am Lehrstuhl entwickelten Peer Data Management System „System P“ werden solche Statistiken in Form von mehrdimensionalen Histogrammen (STHoles [1]) für Kardinalitätsschätzungen innerhalb der Anfragebearbeitung bereits eingesetzt; diese bilden unter Nutzung von query-feedback die Verteilung der Daten über den Wertebereich aller Attribute der Ergebnismenge eines Mappings ab. STHoles-Histogramme erweisen sich als besonders geeignet für die Verwendung im PDMS-Umfeld, da sie den autonomen Charakter des PDMS vollständig erhalten. Es wird angestrebt, diese Histogramme auch auf Paare von Mappings anzuwenden und damit den Überlappungsgrad der hinter diesen Mappings vorhandenen Daten zu beschreiben.

Auf der Grundlage gegebener STHoles-Histogramme für Paare von Mappings und eines im Rahmen der Arbeit zu entwickelnden einfachen Kostenmodells kann dann eine Bewertung alternativer Mapping-Pfade vorgenommen werden bzw. es können möglicherweise sogar komplette Mapping-Pfade bereits bei der Anfrageplanung ausgesondert werden (Pruning).

Statt des Prunings eines Mappings bzw. Mapping-Pfades kann es sich jedoch auch als sinnvoll erweisen, lediglich ein „partielles“ Pruning vorzunehmen. Hierbei werden die umformulierten Anfragen eines Anfrageplanes mithilfe von zusätzlich in die Anfrage eingebrachten Selektionen derartig angepasst, dass Gebiete im multidimensionalen Wertebereich, für die ein hoher Overlap erwartet wird, nicht mehrfach angefragt werden. Diesen Ansatz zu explorieren wird der zentrale Aspekt dieser Arbeit sein.

Hierzu ist es zunächst nötig, einen Algorithmus zu entwickeln, der möglichst große zusammenhängende Bereiche mit (erwartet) hohem Daten-Overlap innerhalb eines gegebenen STHoles-Diagramms identifiziert. Ein erster Ausgangspunkt kann hier möglicherweise die den STHoles-Histogrammen inhärente Baumstruktur sein, mit deren Knoten jeweils eine Histogrammzelle verknüpft ist. Da auch Hose et al. für die in ihrem Ansatz [2] verwendeten „Distributed Data Summaries“ Baumstrukturen nutzen, die der Struktur der STHoles ähnlich sind, ergeben sich hier unter Umständen Ansatzpunkte für den zu entwickelnden Algorithmus. Möglicherweise kann auch auf bekannte Techniken aus dem Bereich der Bäume im Allgemeinen (z.B. R-Bäume, B-Bäume) zurückgegriffen werden.

Die mithilfe eines geeigneten Algorithmus identifizierten Bereiche können nun in einem nächsten Schritt aus der jeweiligen Selektionsanfrage „herausgeschnitten“ und damit zur Abschwächung dieser Anfrage genutzt werden.

Abbildung 1 zeigt ein 2-dimensionales Histogramm für ein Paar alternativer Mappings m_1 und m_2, dass die Verteilung des erwarteten Overlaps O dieser Mappings über den Wertebereich der Attribute x_1 und x_2 abbildet. Auf der Basis dieses Overlap-Histogramms können nun die von einem Peer im Rahmen seiner lokalen Anfragebearbeitung über diese Mappings weiterzuleitenden (umformulierten) Anfragen derart mithilfe von zusätzlichen in die Anfrage eingebrachten Selektionsprädikaten modifiziert werden, dass Gebiete mit zu erwartend hohem Overlap nicht mehrfach angefragt werden.
Die an den über m_1 zu erreichenden Peer 1 gestellte Anfrage Q_1 umfasst also beispielsweise das gesamte Selektionsgebiet der ursprünglichen Anfrage Q, die damit unverändert an den Peer gestellt wird (Q_1 = Q), während aus der über m_2 weitergeleiteten Anfrage an Peer 2 (Q_2 = Q \ O) der Overlap herausgeschnitten wurde. Das Gebiet $Q \cap O$ wird somit nur an Peer 1, nicht aber an Peer 2 angefragt.

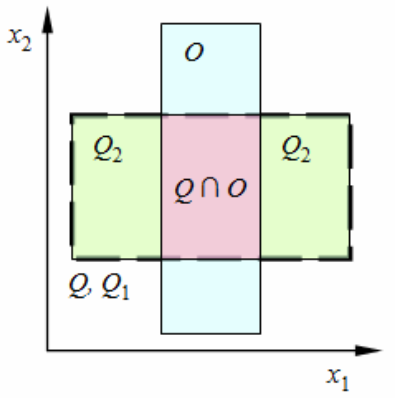

Abb. 1: Anfrage Q, Overlap O und resultierende Anfragen Q_1 und Q_2 an alternative Nachbar-Peers 1 und 2.

Auch hierzu soll im Rahmen der Arbeit eine geeignete Vorgehensweise entwickelt werden, wobei unter Umständen zwischen Feingranularität (d.h. der Anzahl der zusätzlich in die Anfrage eingebrachten Selektionsprädikate) auf der einen Seite und Aspekten wie z.B. Praktikabilität (Ausführbarkeit, Genauigkeit) bzw. Einsparungspotential auf der anderen Seite abgewogen werden muss.

Abschließend sollen unter Berücksichtigung analysierter möglicher Szenarien von Overlap in einem verteilten PDMS die entwickelten Algorithmen im Rahmen der konkreten Implementierung im System P experimentell auf ihre Effektivität und Effizienz untersucht werden.

Literatur

[1] N. Bruno, S. Chaudhuri, and L. Gravano. STHoles: A Multidimensional Workload-Aware Histogram. In *Proc.of the ACM Int. Conf. on Management of Data (SIGMOD)*, 2001.

[2] K. Hose, C. Lemke, and K.-U. Sattler. Processing Relaxed Skylines in PDMS Using Distributed Data Summaries. In *Proc. of the Conf. on Information and Data Management (CIKM)*, 2006.