Bachelorarbeit/Masterarbeit

Karlsruher Institut für Technologie

Institut für Programmstrukturen und Datenorganisation (IPD)
Lehrstuhl für Systeme der Informationsverwaltung, Prof. Böhm

Bachelorarbeit/Masterarbeit

Inhaltliche Analyse von Query-Log Daten einer wissenschaftlichen Datenbank

In dieser Arbeit soll es darum gehen, eine große Menge von SQL-Anfragen, die unterschiedliche Benutzer an eine Datenbank gestellt haben, systematisch zu analysieren, mit den Zielen, besser zu erkennen, wofür die Nutzer sich eigentlich interessieren, und die Performance gesamthaft zu verbessern. Dies soll mit Hilfe von Data-Mining Methoden geschehen (die hier aber in erster Linie nur angewendet werden sollen und voraussichtlich nicht weiterentwickelt werden müssen)**. Konkreter Gegenstand der Betrachtung ist die frei verfügbare SkyServer-Datenbank, eine bekannte Ressource in der Astronomie, für die auch Anfrage-Logs bestehend aus Millionen von Anfragen vorliegen.**

Im Kontext von Datenbanken gibt es diverse Performance Antipatterns, d. h. wiederholt auftretende, ineffiziente Darstellungen von Informationsbedürfnissen in SQL. Eine erste Analyse, die wir selbst durchgeführt haben, hat ergeben, dass manche Antipatterns im SkyServer Anfrage-Log häufig vorkommen. Diese Antipatterns machen nicht nur die Anfrageauswertung langsamer, sie erschweren auch das o. g. Erkennen von Benutzerinteressen auf inhaltlicher Ebene. Ein Antipattern ist beispielsweise, eine sinnvolle SQL-Anfrage unnötigerweise in mehrere Anfragen zu zerlegen und den Join dieser Anfragen dann in die Anwendung zu verlagern: Eine erste Anfrage könnte beispielsweise IDs von Datenobjekten zurückliefern, die in einer zweiten Anfrage Teil der WHERE-Klausel sind. Unsere Hypothese ist, dass herkömmliche Techniken, die aufbauend auf dem Anfrage-Log Benutzern Anfragen empfehlen, oder die herausfinden, welche Teile des Datenraums viele Benutzer gemäß ihrer Anfragen interessant finden, in Gegenwart solcher Antipatterns nicht gut funktionieren. Diese Aufgabe umfasst im Wesentlichen die folgenden Punkte:

- Herleiten von Aussagen dazu, wie häufig bestimmte Antipatterns im SkyServer Anfrage-Log vorkommen. Wir haben bereits damit begonnen, SkyServer-Queries zu analysieren, und Sie können von uns entwickelte Werkzeuge wiederverwenden.
- Entwicklung von Techniken, die Antipatterns aus dem Anfrage-Log 'entfernen', indem sie entsprechende Anfragen durch gleichbedeutende, aber bessere Anfragen ersetzen.
- Durchführung von Untersuchungen, wie sehr bestimmte Antipatterns jene inhaltlichen Analysen des Anfrage-Logs verfälschen. Dies, indem die Analyseergebnisse auf dem ursprünglichen Anfrage-Log mit Analyseergebnissen auf einem 'bereinigten' Anfrage-Log (d. h. jene Antipatterns wurden durch bessere Queries ersetzt) verglichen werden. Auch für diese Analyse steht Code zur Verfügung, den Sie grundsätzlich verwenden können.

Sie erwerben durch die Bearbeitung dieser Aufgabe eine hohe Vertrautheit sowohl mit Datenbanktechnologie, die in dieser Form auch in Firmen und anderen Organisationen zum Einsatz kommt, als auch mit modernen Methoden der Datenanalyse. Dies sowohl auf der konzeptionellen Ebene als auch, was den konkreten praktischen Einsatz entsprechender Technologie angeht. – Der genaue Umfang der Arbeit lässt sich variieren, dadurch ist die Bearbeitung als Bachelor- als auch als Masterarbeit möglich.

Ansprechpartner

Prof. Dr. Klemens Böhm klemens.boehm@kit.edu +49 721 608-43968 Raum: 366

Am Fasanengarten 5 76131 Karlsruhe Gebäude: 50.34