Karlsruher Institut für Technologie

Institut für Angewandte Informatik und Formale Beschreibungsverfahren

Graduiertenkolloquium Angewandte Informatik

Integration von multidimensionalen Datensätzen aus dem Web für analytische Abfragen

Dipl.-Inform. Benedikt Kämpgen
AIFB

In verschiedenen Bereichen werden zunehmend Statistiken oder Sensordaten im Web veröffentlicht. Die Integration solcher multidimensionalen Datensätze verspricht einen hohen Mehrwert. So können Naturwissenschaftler weltweit gesammelte Klimadaten mit Analyseergebnissen aus Publikationen ergänzen. Finanzanalysten können Kennzahlen aus den Jahres- und Quartalsberichten in Bezug zu Börsendaten setzen, um amerikanische Firmen zu bewerten. Und europäische Bürger können statistische Indikatoren wie das Bruttoinlandsprodukt pro Kopf anhand von Daten verschiedener Institutionen, z.B. Eurostat, für alle Staaten berechnen und vergleichen.

Obwohl die Datensätze wohlstrukturiert im Web verfügbar sind, ist deren Integration und Analyse weiterhin problematisch. Denn Datensätze sind heterogen, weisen z.B. unterschiedliche Bezeichner für Dimensionen wie "Geo" und "Ort" oder Dimensionswerte wie "DE" und "Deutschland" auf. Desweiteren erfordert die Anzahl und Größe der Datensätze die Optimierung von Abfragen; allein Eurostat veröffentlicht mehr als 5,000 Datensätze.

In der vorgestellten Arbeit ermöglichen wir die Integration von Datensätzen auch bei impliziten Überlappungen durch die Drill-Across-Operation. Für die Integration bei komplexeren Beziehungen zwischen Datensätzen stellen wir Konvertierungs- und Kombinationsoperationen vor. Eine Skalierbarkeit von Abfragen erreichen wir durch die Wiederverwendung bestehender OLAP-Server. Analytische Abfragen direkt auf Datensätze in einer RDF-Datenbank optimieren wir durch die Materialisierung von Aggregationen.

Als Grundlage mappen wir zwischen dem bekannten Multidimensionalen Datenmodel für Datenwürfel und einem weitverbreiteten Vokabular zur Veröffentlichung von multidimensionalen Datensätzen als Linked Data. Außerdem übersetzen wir bekannte OLAP-Operationen wie Slice und Dice in Abfragen über Datensätze in einer RDF-Datenbank. Wir untersuchen die Anwendbarkeit der Ansätze auf Szenarien in den Naturwissenschaften, der Finanzanalyse und der Verwendung öffentlicher Statistiken.

Termin: Mittwoch, 11. Juni 2014, 14.00 Uhr

Ort: Englerstraße 11, 76131 Karlsruhe
Kollegiengebäude am Ehrenhof (Geb. 11.40), 2. OG, Raum 253
(Hinweise für Besucher: www.aifb.kit.edu/web/Kontakt)

Veranstalter: Institut AIFB, Forschungsgruppe Wissensmanagement

Zu diesem Vortrag lädt das Institut für Angewandte Informatik und Formale Beschreibungsverfahren alle Interessierten herzlich ein.

Andreas Oberweis, Hartmut Schmeck, Detlef Seese, Wolffried Stucky, Rudi Studer (Org.), Stefan Tai

Besucheranschrift:
KIT-Campus Süd
Institut AIFB – Geb. 11.40
Kaiserstr. 89
76133 Karlsruhe

Postanschrift:
KIT-Campus Süd
Institut AIFB – Geb. 11.40
76128 Karlsruhe