Bachelorprojekt

Wintersemester 2016/2017 – Sommersemester 2017

Graph Partitionierung für Graphdatenbanken

Projekthintergrund

Graphdatenbanken gewinnen, unter anderem für Anwendungen im Bereich von sozialen, wissenschaftlichen und Geschäftsnetzwerken, in der Softwaretechnik und in biologischen Netzwerken, an zunehmender Bedeutung. Im Gegensatz zu relationalen Datenbanken unterliegen die Inhalte von graphbasierten Datenbanken einer Graphtopologie und benutzen spezialisierte Algorithmen, um Suchanfragen effizient zu bearbeiten. Die Suchanfragen können z.B. über Grapherweiterungen für SQL modelliert werden und müssen durch geeignete Graphalgorithmen implementiert werden, um eine hohe Performanz zu erreichen.

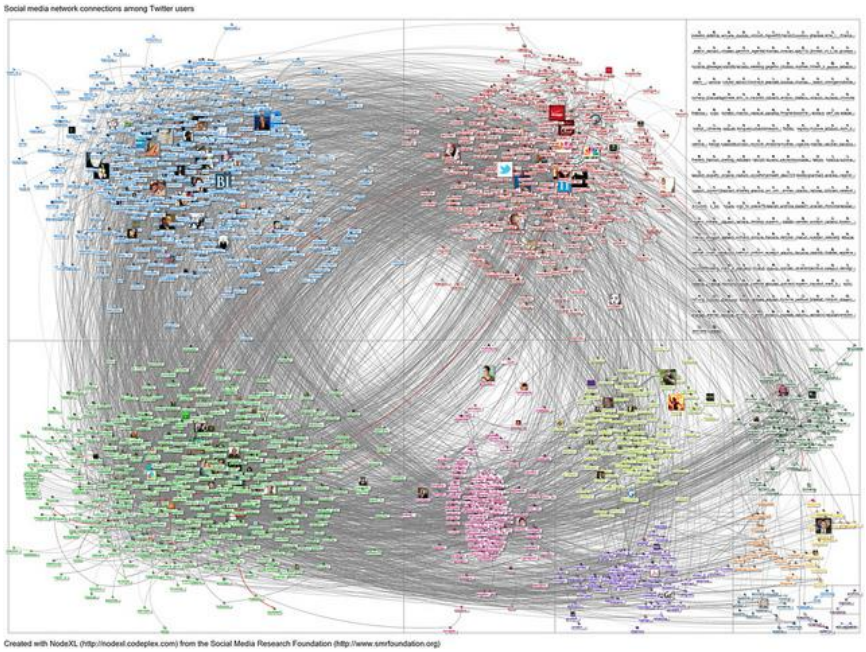

Abbildung 1: Ausschnitt aus dem Twitter Netzwerk für einen Nutzer[1]

Partitionierung von Graphen ist eine Technik um den Graphen in unterschiedliche Bereiche (Partitionen) zu teilen. Der Nutzen einer solchen Partitionierung ist zum einen die Möglichkeit Suchanfragen zu parallelisieren, wobei die Partitionierung vorgibt in welchen Teilen des Graphen parallel gesucht wird. Dadurch ermöglichen Partitionierungen eine deutliche Steigerung der Performanz. Zum anderen ist Partitionierung nötig um sehr große Graphen (Milliarden Knoten, wie z.B. das gesamte Twitter Netzwerk, Ausschnitt siehe Abb. 1) verteilt zu analysieren. Hierbei bestimmt die Partitionierung wie Graphdaten in einem Serververbund verteilt gespeichert und analysiert werden.

In diesem Bachelorprojekt sollen zusammen mit dem SAP Innovation Center basierend auf bestehenden Partitionierungsalgorithmen für Graphen Erweiterungen einer sich bei SAP in der Entwicklung befindlichen verteilten Graphdatenbank-Software für SAP HANA Vora[2] konzipiert und prototypisch implementiert werden. Insbesondere soll der Einfluss der Partitionierung auf die Auswertungseffizienz der Suchanfragen untersucht werden. Dabei spielen eine flexible Integration unterschiedlicher Partitionsstrategien in die bestehende Graphdatenbank und deren systematischen Evaluierung eine Hauptrolle.

---

[1] http://theincidentaleconomist.com/wordpress/wp-content/uploads/2014/06/NodeXL-Twitter-User-jowyang-network-graph.jpg

[2] http://go.sap.com/product/data-mgmt/hana-vora-hadoop.html

Projektgegenstand

Ziel des Bachelorprojektes ist die Konzeption und Implementierung von mehreren Erweiterungen einer sich bei SAP in der Entwicklung befindlichen verteilten Graphdatenbank und deren Evaluierung. Als erste Erweiterung soll ein Mechanismus entwickelt werden, der es erlaubt existierende Partitionierungsstrategien flexibel in die Graphdatenbank zu integrieren. Dazu muss der bisherige grundlegende Partitionierungsmechanismus verstanden und erweitert werden. Als zweite Erweiterung soll ein nach der Partitionierungsstrategie parametrisiertes Benchmarkingframework entwickelt werden, das auf existierende Frameworks für die Auswertung von Suchanfragen aufbaut. Mit Hilfe dieser Erweiterungen sollen die Auswirkungen unterschiedlicher Partitionierungsstrategien auf die Effizienz der Auswertung der verschiedenen Suchanfragen systematisch evaluiert werden. Existierende Graphvisualisierungswerkzeuge, wie z.B. Gephi (https://gephi.org/), können hierbei benutzt werden um verschiedene Partitionierungsstrategien graphisch zu analysieren und zu vergleichen. Mit dieser Evaluierung als Grundlage soll eine Heuristik konzipiert und implementiert werden, die eine möglichst optimale Partitionierungsstrategie dynamisch wählt. Schlussendlich sollen im Ausblick Kernideen für neuartige Anforderungen an Partitionierungsalgorithmen erarbeitet werden, die die bei der Evaluierung herausgefundenen Schwächen ausgleichen.

Umsetzung

Die ersten Schritte im Bachelorprojekt umfassen die Einarbeitung in die Themen Graphdatenbanken und Partitionierungsalgorithmen für Graphen sowie relevante Dokumentation. Desweiteren wird eine Einführung in die bei SAP befindliche Datenbankarchitektur gegeben und sowohl formale als auch technische Grundlagen für die Erweiterungen gegeben. Nach einer detaillierten Anforderungsanalyse werden für die zu entwickelnden Teile passende Designs entwickelt und prototypisch implementiert. Abschließend werden die jeweiligen Erweiterungen mit Hilfe des Benchmarkingframeworks evaluiert.

Projektumfeld

Das Projekt findet in Zusammenarbeit mit dem SAP Innovation Center in Potsdam statt. Dort befindet sich eine Graphdatenbank in Entwicklung welche als Grundlage für die zu entwickelnden Erweiterungen dient.

Organisation

In der Seminarphase werden durch die teilnehmenden Studenten die Grundlagen zur genutzten Graphdatenbank-Software und zu ausgewählten Themen erarbeitet und präsentiert (1. Meilenstein). Die Ergebnisse der Seminarphase bilden die Grundlage für die folgende Anforderungserhebung, die in einem Anforderungsdokument zusammengetragen wird (2. Meilenstein). Das Anforderungsdokument wird in einem Antrittsvortrag dem SAP Innovation Center vorgestellt. Auf Basis der Anforderungen werden dann entsprechende Konzepte erarbeitet, die in einem Designdokument beschrieben werden (3. Meilenstein). Die Umsetzung der Konzepte wird in Form der Bachelorarbeiten beschrieben und evaluiert (4. Meilenstein). Abschließend werden die Ergebnisse des Bachelorprojektes in Form eines Abschlussvortrags vor dem SAP Innovation Center präsentiert (5. Meilenstein). Bei gutem Gelingen wird eine Anstellung als studentische Hilfskraft in Aussicht gestellt.

Teilnehmer und Projektbeginn

Bis zu 6 Teilnehmer können in diesem Projekt mitarbeiten. Projektbeginn ist der 1.10.2016.

Informationen

Für ausführliche Informationen zu dem Projekt stehen Prof. Holger Giese (holger.giese@hpi.de) und Senior Researcher Dr. Leen Lambers (leen.lambers@hpi.de) zur Verfügung. Ansprechpartner seitens des SAP Innovation Centers sind Dr. Christian Krause (christian.krause01@sap.com) und Dr. Daniel Johannsen (daniel.johannsen@sap.com).

SAP Innovation Center

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam