Karlsruher Institut für Technologie

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

Diplomarbeit / Masterarbeit

Frequent-Subgraph-Mining unter Nutzung von Hierarchien

Ein wichtiges Problem im Data Mining ist das Finden von Frequent Itemsets (wie aus der Vorlesung Data Warehousing und Mining bekannt). Das Problem, häufige Strukturen zu finden, lässt sich für Graphen verallgemeinern. Sogenannte *Frequent-Subgraph-Mining-Algorithmen* finden Anwendung in zahlreichen Gebieten. Diese Algorithmen suchen in einer Menge von Graphen nach Teilgraphen, die mit einer gewissen Mindesthäufigkeit (*Support*) auftreten. Um häufige Teilgraphen zu finden, generiert man zunächst Kandidaten und ermittelt anschließend deren Häufigkeit. Dazu ist ein Test auf Subgraph-Isomorphie nötig. Dieser Test ist NP-vollständig. Frequent-Subgraph-Mining-Algorithmen skalieren aus diesem Grund schlecht.

Es existieren verschiedene Ansätze, die durch Beschneidung (*Pruning*) des Suchraumes die Kandidatenzahl reduzieren und somit auch die Anzahl der Subgraph-Isomorphie-Tests. Dadurch kann eine signifikante Verbesserung der Laufzeit erzielt werden. In dieser Arbeit soll ein Konzept zur Beschneidung des Suchraums erarbeitet und evaluiert werden, das orthogonal zu bisherigen Konzepten ist, wie folgt: Häufig haben Graphen eine gegebene Hierarchie. Beispielsweise ließen sich bei einer Straßenkarte (Kreuzungen sind Knoten, Straßen Kanten) Kreuzungen zu Städten, Ländern, etc. zusammenfassen. Die Größe des Graphen nimmt mit sinkender Granularität ab. Eine Analyse benötigt daher entsprechend wenig Rechenzeit. Die Informationen, die auf aggregierten Graphen gewonnen werden, können bei der Analyse auf dem ursprünglichen Graphen (oder weniger stark aggregierten Graphen) genutzt werden. Stellt man beispielsweise fest, dass es keinen häufigen Graphen mit der Kante *Deutschland->Japan* gibt, brauchen Kandidaten mit einer Kante zwischen München und Tokyo nicht betrachtet werden. Sollte keine natürliche Hierarchie für einen Graphen vorliegen, kann eine solche auch künstlich erzeugt werden.

Im Rahmen dieser Arbeit soll eine vorhandene Implementierung von gSpan [1] erweitert werden, so dass das beschriebene Pruning-Kriterium genutzt werden kann. Der erweiterte Algorithmus soll in einer Evaluation mit der ursprünglichen Version verglichen werden.

Bei der Bearbeitung der Aufgabe lässt sich Erfahrung mit der Analyse komplexer Datenbestände sammeln. Es werden aktuelle Algorithmen eingesetzt und es kann Beitrag zu einem wichtigen Gebiet im Data Mining geleistet werden.

[1] Yan, X. & Han, J.; gSpan: Graph-Based Substructure Pattern Mining; Department of Computer Science, University of Illinois at Urbana-Champaign, 2002

Ansprechpartner

Dipl. Inform. Christopher Oßner ossner@kit.edu +49 721 608 44065 Raum: 340

Am Fasanengarten 5 76131 Karlsruhe Gebäude: 50.34

Diplomarbeit / Masterarbeit