



Diplomarbeit

Implementieren des Subspace Clustering Algorithmus

Der Subspace Clustering Algorithmus ermittelt neben den Clustern, die auch jedes beliebige andere Clusterverfahren errechnet, noch eine Projektion für jeden Cluster. Clusterverfahren teilen automatisch Objekte in unterschiedliche Gruppen ein. Die Objekte einer Gruppe sind sich nach einem gegebenen Maß ähnlich und unterscheiden sich möglichst stark von Objekten der anderen Gruppen.

Beim Clustern im hochdimensionalen Raum führen verschiedene Maße zur Berechnung der Ähnlichkeit zu Problemen. Eine Projektion, die die Anzahl der Dimensionen reduziert führt typischerweise zu einem hohen Informationsverlust. Die Idee des Algorithmus ist es die Projektion nicht für den gesamten Raum, sondern für jeden Cluster separat durchzuführen.

Die Aufgabe der Diplomarbeit ist es, den Subspace Clustering Algorithmus in Java effizient zu implementieren und an die Datenbank DB2 anzuschließen. Anschließend soll der Algorithmus so erweitert werden, dass Hintergrundwissen in den Projektionsprozess einfließt. In Tests ist die Effizienz der Implementierung sowie die Verbesserung der Güte in Bezug auf die Clusterergebnisse zu zeigen.

Literatur

- o R. Agrawal, J. Gehrke, D. Gunopulos, P. Raghavan: "Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications", Proc. of the ACM SIGMOD Int'l Conference on Management of Data, Seattle, Washington, June 1998.
- o Hotho, A., Mädche, A., Staab, S.: Ontology-based Text Clustering, Workshop "Text Learning: Beyond Supervision", IJCAI 2001.

Bei Interesse melden sie sich bitte bei:

Andreas Hotho

Raum: 252

Tel.: 0721-608-6558

E-Mail: hotho@aifb.uni-karlsruhe.de

Institut (AIFB), Universität Karlsruhe (TH)

76128 Karlsruhe

Fax: 0721-693717