



Diplomarbeit

Vergleich LSI, Concept Indexing und COSA

Mit Methoden und Verfahren zur Extraktion von Daten aus Texten beschäftigt sich das **Information Retrieval**. Dazu ist es nötig, Texte in einer geeigneten Art und Weise zu repräsentieren. Texte können unter anderem durch einen sogenannten Wortvektor repräsentiert werden, wobei die Häufigkeit eines jeden Wortes im Text gezählt wird. Da die natürliche Sprache viele unterschiedlichen Worte aufweist, führt diese Repräsentation zu Vektoren mit vielen Elementen, der Vektor spannt also einem hochdimensionalen Raum auf. Diese hohe Anzahl an Dimensionen führt aber bei Verfahren wie dem Clustering zu Problemen.

Die Verfahren LSI, Concept Indexing und COSA werden zur Dimensionsreduktion eingesetzt, wobei die Ergebnisse bei COSA trotz Dimensionsreduktion interpretierbar bleiben. Die Verfahren führen eine Datenvorverarbeitung durch und führen so bei nachgeschalteten Algorithmen (z.B. Clusterverfahren) zu besseren Ergebnissen. COSA bietet zusätzlich den Vorteil, das auch der Mensch in der Lage ist, die niedrigdimensionalen Daten zu verstehen und so die erzeugten Ergebnisse zu interpretieren.

Die Aufgabe der Diplomarbeit ist es, anhand verschiedener Datensätze die Güte der verschiedenen Verfahren als Vorverarbeitungsschritt zu vergleichen.

Literatur

- o Hotho, A., Mädche, A., Staab, S.: Ontology-based Text Clustering, Workshop "Text Learning: Beyond Supervision", IJCAI 2001.
- o George Karypis, Eui-Hong (Sam) Han, Concept Indexing A Fast Dimensionality Reduction Algorithm with Applications to Document Retrieval & Categorization (2000), Technical Report TR-00-0016, University of Minnesota, 2000.
- o S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R. Hashman. Indexing by latent semantic indexing. Journal of the American Society for Information Science, 41(6),1990.
- o Blake, C.L. & Merz, C.J. (1998). UCI Repository of machine learning databases [<http://www.ics.uci.edu/~mlearn/MLRepository.html>]. Irvine, CA: University of California, Department of Information and Computer Science.
- o The Reuters-21578 Text Categorization Test Collection <http://www.research.att.com/~lewis/reuters21578.html>

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