Prof. Dr.-Ing. Thomas Kudraß
HTWK Leipzig, FB IMN

Datenbanken-Themen im OS "Data Mining" WS 2008/2009

Nachfolgend finden Sie einige Hinweise und Fragestellungen zu den ausgegebenen Themen.

Methodik des Data Mining

1. KDD-Prozess

Es sind die Phasen des KDD-Prozesses (Knowledge Discovery in Databases) zu beschreiben u ihre jeweiligen Besonderheiten: Dazu zählen im Einzelnen:

- Datenselektion und -extraktion
- Datenaufbereitung (Vorverarbeitung)
- Datenbereinigung und Transformationen (Aggregation, Berechnung, korrelierte Attribute, Transsformation, Codierung, Normierung) und Anwendungen
- Data Mining
- Interpretation -Quellen:

Verfahren und Algorithmen des Data Minig

2. Clustering

Beim Clustering werden Datenobjekte in Klassen ähnlicher Objekte zusammengefasst, d.h. in sogenannte Cluster. In einem Cluster sind die Objekte zueinander möglichst ähnlich und zu Objekten anderer Cluster möglichst unähnlich. Der Vortrag sollte auf folgende Aspekte eingehen:

- Anwendungen für Clustering
- Anforderungen und zu lösende Probleme
- Überblick über Clustering-Verfahren (insb. Datenunabhängigkeit)
- Vorstellung von Algorithmen im Detail (mindestens zwei): k-means / EM, ein hierarchisches Verfahren

3. Assoziationsanalyse

Bei der Assoziationsanalyse geht es um das Finden von Regeln, die das Auftreten eine Items in Abhängigkeit vom Auftreten anderer Items vorhersagen. Der Vortrag sollte auf folgende Aspekte eingehen:

- Anwendungen für Assoziationsanalyse (Warenkorb)
- Verfahren: Brute-Force-Algorithmus, A-Priori-Algorithmus, FP-Growth-Algorithmus
- Finden von mehrdimensionalen Assoziationsregeln aus relationalen Datenbanken und Data Warehouses

- Ausblick: Von der Assoziationsanalyse zur Korrelationsanalyse

4. Klassifikation (1-2 Vorträge)

Die Klassifikation ist das Zuordnen von Datenobjekten zu einer Klasse aus einer vorgegebenen Menge von Klassen. Der Vortrag sollte folgende Aspekte beinhalten:

- Anwendungen
- Prinzip (Ermittlung Klassifikator, Zuordnung von Klassen) + Anforderungen
- Klassifikationsalgorithmen:
- Entscheidungsbaum-Klassifikatoren
- Bayes-Verfahren
- Künstliche Neuronale Netzwerke / Klassifikation durch Backpropagation
- Weitere Verfahren: k-Nearest-Neighbour-Klassifikator, Support Vector Machines (SVM)
- Abstands- und Ähnlichkeitsmaße
- Dimensionsreduktion (Feature Selection, Principal Component Analysis)

5. Anomalieentdeckung

Bei der Anomalientdeckung werden Ausreißer aufgefunden, die dadurch definiert sind, dass sie offensichtlich anders sind als die anderen Objekte der Datenmenge. Der Vortrag sollte auf folgende Aspekte eingehen:

- Anwendungen
- Anforderungen und Probleme
- Überblick über Verfahren zur Anomalieentdeckung
- Beispielhafte Präsentation zweier Verfahren (grafik- oder statistikbasiert, distanzbasiert)

Systeme und Tools zum Data Mining

6. RapidMiner als Beispiel für Open-Source-Software

RapidMiner (zuvor YALE) ist eine universitär entwickelte Open-Source-Software unter der AGPL und beinhaltet ca. 400 Operatoren zur Wissensextraktion in Datenbanken (KDD). Der Vortrag sollte folgende Aspekte beinhalten:

- Überblick über die Funktionalität von RapidMiner: Schnittstellen, Oberfläche, Visualisierung, Erweiterungsmöglichkeiten, Datenformate
- Kurze Produktpräsentation

7. Data Mining mit den Microsoft SQL-Server 2005 Analysis Services

Data-Mining-Ansätze haben schon seit einigen Jahren in kommerziellen Datenbankprodukten Einzug gehalten. Stellvertretend hierfür sollen die Möglichkeiten des SQL Server 2005 vorgestellt werden. Der Vortrag sollte auf folgende Aspekte eingehen:

- unterstützte Data-Mining-Algorithmen

- DMX - Data Mining Query Language
- Architektur der Data-Mining-Komponente und System-Schnittstellen (XML for Analysis API, Erweiterungsmöglichkeiten)

Spezielle Anwendungen des Data Mining

8. Text Mining

Text Mining bezeichnet den Prozess der Wissensentdeckung in textuellen Daten. Texte sind im Gegenstand zu relationalen Daten, die den Gegenstand des Data Mining bilden, unstrukturierte oder schwach strukturierte Daten. Entsprechend ist zunächst eine strukturelle und linguistische Analyse erforderlich, um die Daten in eine strukturierte Form zu bringen, die die Anwendung von Methoden des Data Mining ermöglicht. Der Vortrag sollte auf folgende Aspekte eingehen:

- Anwendungen des Text Mining
- Vorverarbeitung / Linguistische Analyse von Texten
- Verfahren des Text Mining: Grundlagen (Vektorraum-Modell, Relevanzbewertung von Termen), Assoziationsanalyse innerhalb von Texten, Analyse von Beziehungen zwischen Informationen aus verschiedenen Texten

9. Web Mining

Web Mining beinhaltet Techniken aus dem Data Mining zur automatischen Extraktion von Informationen aus dem Internet (WWW). Dabei unterscheidet man drei Arten des Web Mining:

- Web Content Mining: Anwendung des Text Mining auf Webseiten
- Web Structure Mining: Analyse der Verweisstrukturen (Hyperlinks) einer Webseite. Dies kann zur Kategorisierung oder zum Ranking von Webseiten genutzt werden.
- Web Usage Mining: Analyse des Benutzerverhaltens im Web (z.B. auf der Basis von Logdateien). Dies kann zur Erstellung von Benutzerprofilen genutzt werden und damit zur Personalisierung von Internetpräsenzen.

Der Vortrag sollte einen Schwerpunkt auf Web Usage Mining setzen und dabei auf auf folgende Aspekte eingehen:

- Anforderungen an Web-Zugriffsanalysen
- Realisierung technischer Lösungen
- KDD-Prozess
- Schnittstellen zu Datenbank / Data Warehouse
- Tools

Quelle:

- Rahm, E., Stöhr, T. Data-Warehouse-Einsatz zur Web-Zugriffsanalyse, in: Rahm, E.; Vossen, G.: (Hrsg.): *Web und Datenbanken*. dpunkt-Verlag, 2003.

10.Spatial Data Mining

Spatial Data Mining ist die Anwendung von Data-Mining-Techniken auf Geodaten (räumliche Daten). Fast alle in Datenbanken und Data Warehouses gespeicherten Daten haben einen Raumbezug. Hierzu tragen technologische Entwicklungen wie GPS, RFID und Sensornetze bei. Im Vortrag sind die Besonderheiten von raum-zeitlichen Daten gegenüber herkömmlichen Daten herauszuarbeiten und die besonderen Analysetechniken des Spatial Data Mining darzustellen. Der Vortrag sollte auf folgende Aspekte eingehen:

- Charakteristik von Geodaten
- Anwendungen, z.B. Geomarketing, Reichweitenbestimmung von Werbemedien, Wirtschaftsgeographie
- Ausgewählte Algorithmen des Spatial Data Mining mit den jeweiligen Anwendungen, z.B: Spatial Clustering, Spatial Association Rules, Spatial Classification

Gesellschaftliche Aspekte des Data Mining

11.Data Mining und Datenschutz

Data Mining eröffnet vielfältige technische Möglichkeiten, die von Datenschützern kritisch beäugt werden. Die Metapher "gläserner Mensch" beschreibt die Befürchtung, dass wir immer mehr und überall durchleuchtbar werden.

Der Vortrag sollte folgende Aspekte beinhalten, kann aber gern um aktuell interessierende Themen erweitert werden.

- Datenschutzrechtliche Bestimmung vs. Data Mining (Konflikte?)
- Staatliche Projekte (MATRIX u.ä.)
- Möglichkeiten aus neuen Technologien (RFID)
- Web 2.0, Data Mining und der Schutz persönlicher Daten im Internet

Quellen für alle Vorträge (Einstiegsliteratur):

- Witten, Ian H.; Frank, E.: *Data Mining. Praktische Werkzeuge und Techniken für das maschinelle Lernen.* Hanser-Verlag, 2001.
- Han, J.; Kamber, M.: *Data Mining. Concepts and Techniques*. Morgan Kaufmann Publ., 2000.
- Herden, O.: Data Mining. in: Kudraß, T. (Hrsg.): *Taschenbuch Datenbanken*. Hanser-Verlag, 2007.

Graduierungsarbeiten auf dem Gebiet Data Mining

- Bärthel, D.: *Automatische Textklassifikation. Erweiterung des DMS der forcont factory um eine automatische Dokumentarterkennung*. Diplomarbeit HTWK Leipzig 2008.
- Kunze, A.: *Data-Mining Verfahren zur automatischen Bedarfsermittlung im Rahmen von IT-Weiterbildung*. Masterarbeit HTWK Leipzig 2005.