Prüfungsprotokoll zur Vertiefungsprüfung Datenmanagement und Datenexploration

an der RWTH Aachen

Prüfer: Prof. Seidl
Vorlesungen:

- Einführung in Datenbanken (nach Buch „Datenbanksysteme“, Kemper/Eickler)
- Data Mining Algorithms (nach Folienskript SS05, Prof. Seidl)
- Indexstrukturen (nach Folienskript WS05/06, Prof. Seidl)

Datum: 22.03.2006
Dauer: 45 Minuten
Note: 1.0

Bemerkung: Die Atmosphäre war sehr freundlich und entspannt. Mit AOI konnte ich nichts anfangen, wir haben es dann zusammen „erarbeitet“, im Endeffekt war das aber nicht schlimm. Mein Tip: Prof. Seidl auch mal ein Wenig erzählen lassen.

Einführung in Datenbanken:

- Welche Grundlegenden Datenbankmodelle gibt es?
- Was ist das Besondere an OO Datenbanken?
- Was sind die Schwachpunkte der Relationalen Modellierung?
- Was ist eine Algebra ganz allgemein?
    - Worauf arbeitet eine Algebra?
    - Worauf arbeitet die Relationale Algebra?
- Was beinhaltet die RA noch? Schreiben Sie mal die Operationen auf.
- Was gibt’s noch? (TK, DK → Aussagefähigkeit, Turing Vollständigkeit)
- Welche Arten des Joins gibt es?
    - Formulieren Sie einen Join von R = [A, B] und S = [B, C] in SQL, RA, TK, und DK.

Data Mining Algorithms:

- Überleitungsfrage: Wie werden die Daten in einem Data Warehouse gespeichert? Was ist ein Data Warehouse?
    - Welche Werte speichert ein Data Cube?
    - Sternschema, Snowflake Schema erklären.
    - Was sind die typischen Anfragen an ein DW?
    - Drill down, roll up, slice und dice in SQL.
- Wie funktioniert Attribute Oriented Induction (AOI)?
- Was fällt Ihnen generell zum Thema Generalisierung und Konzept Hierarchien ein?
    - Wie verfährt man mit kategorischen Attributen?
    - Wie verfährt man mir reellwertigen Attributen?
    - Was würden Sie bei einer Generalisierung mit einem Attribut „Matrikelnummer“ machen? (→ weglassen)
- Was sind die Aufgaben bei der Klassifikation? Wie grenzt sie sich ab?
- Erläutern Sie den Bayes Klassifikator. (Über Hypothesen hin zum optimalen Bayes Klassifikator.)
    - Schreiben Sie mal die Formeln auf. Warum kann man das P(o) weglassen? (Da P(o) von C_j stochastisch unabhängig ist und *Zitat:“ …es das arg nicht überleben würde.“*)
    - Wieso formt man die Gleichung überhaupt um?
    - Wie kommt man denn an die Werte für P(c) und P(o|c)? (→ naiver Bayes Klassifikator)
    - Wie verfährt man bei kategorischen/reellwertigen Attributen?
    - Ist man ohne die Annahme der stochastischen Unabhängigkeit denn verloren? (→ Nein: Multivariate W’keitsverteilungen. Als Bsp. Gaussformel aufgeschrieben.)
- Was ist der NN-Klassifikator? (Standard und beide Gewichtungsvarianten mit Bsp.)

Indexstrukturen:

- Braucht man Indexstrukturen überhaupt?
- Welche Verfahren kennen Sie zum speichern von eindimensionalen Daten?
- Welche Varianten gibt es beim Hashing?
- Wie geht man beim Bitmap Index vor?
    - Für welche Wertebereiche ist Bitmap Indexing? Was geschieht mit rellen Attributen?
    - Nehmen wir an ein Attribut A hätte drei verschiedene Werte, wie viele Bitmaps benötige ich beim Standard BI? (→ 3) Was gibt es für Alternativen? (→ Komprimierung) Welche Vor- und Nachteile birgt dies?
- Kann ein B-Baum degenrieren? Warum nicht?
- Kann man zur Speicherung von Intervalldaten auch Punktzugriffsstrukturen verwenden?
    - Prinzip der Punkttransformation erläutern und einige Beispielanfragen zur Eckentransformation durchführen.
    - Wie funktioniert MAP21?
- Wie sieht ein R-Baum aus?
    - Kann man darin auch Punktdaten speichern?
    - Erläutern Sie das Einfügen in einem R-Baum.
    - Lassen sich Überschneidungen verhindern?
    - Was sind die Probleme bei hochdimensionalen Daten?