# Gedächtnisprotokoll

## Diplomprüfung Informatik Vertiefung

| | |
|---|---|
| **Datum** | 02.07.2004 |
| **Themen**: | Data Mining Algorithmen (Prof. Seidl, WS 03/04)<br>Indexstrukturen für Datenbanken (Prof. Seidl, SS 04)<br>Implementierung von Datenbanken (nach Härder/Rahm, s. I9 Webseite) |
| **Prüfer**: | Prof. Seidl |

## Data Mining Algorithmen **(ca. 15 Minuten)**

- Welche Kategorien von DM Techniken gibt es? [Clustering, Assoc. Rules, …]
- Optimal Bayes Classifier erklären (mit Formeln)
- Wie bestimmt man die P(x|C_i)? [Ich wollte auf Naiven Bayes Classifier heraus, er wollte Approximation mit d-dim Normalverteilung hoeren]
- Welche Parameter der Normalverteilung muss man dazu schätzen? [d-dim mean und d-dim Kovarianz Matrix]
- Nearest Neighbor Classifier erklären, Beispiel bringen, Varianten aufzaehlen [k-NN, k-NN mit verschiedenen Gewichtungen [nach Distanz und relativer Klassenhäufigkeit]]
- Decision Trees erklären, Overfitting, Pruning (nur die grobe Idee)

## Implementierung von DB **(ca. 15 Minuten)**

- Was sind Transaktionen?
- ACID Eigenschaften genau erklären
- Mehrbenutzeranomalien erklären
- Wie stellt man Isolation her? [Synchronisation]
- Was ist die Idee der Synchronisation? [Serialisierbarkeits Kriterium erklärt]
- Fundamentalsatz der Synchronisation
- Welche Verfahren gibt es [2PL, Hierarchische Erweiterung, optimistisch, …]
- 2 Phase Locking, strict 2PL, strict 2PL mit preclaiming
- Optimistische Verfahren (Idee, BOCC, FOCC erklären). Härder schreibt, dass manchmal für optimistische Verfahren alternativ preclaiming angeboten wird (hybrid), damit einmal zurückgesetzte Transaktionen im nächsten Durchlauf durchkommen (die benötigten DB Objekte sind ja aus dem letzten Lauf bekannt). Hat zur Diskussion geführt: Wir sind zu dem Schluss gekommen, dass sich preclaiming wirklich nur ausserhalb des DB Bereich (C/S Systeme, insb. CVS) durchgesetzt hat.
- Unterschied algebraische/nicht-algebraische Optimierung und Erklärung der beiden
- Kostenmodell in der nicht-algebraischen Optimierung.
- Wie macht man Selektivitätsabschätzungen [ohne Formeln, Idee und Realisierung in DBMS [über periodisch aktualisierte Statistiken für Indexe u.A.]

## Indexstrukturen **(ca. 15 Minuten)**

- Wozu braucht man Indexstrukturen? Was machen sie? Kann man DBMS auch ohne sie implementieren? [Ja klar]
- Hashing: Idee und Eigenschaften von Hashfunktionen, offenes und geschlossenes Hashing, Sondierungsverfahren
- Hashing mit Directory
- B-Bäume: Aufbau, Funktion des Knotensplit, Suche
- R-Bäume: Aufbau, Funktion von MURs
- Nächste Nachbar Anfragen auf R-Bäumen (Naiv, nach Rossopulous und nach Samet)
- Zum Samet Verfahren: Was ist der Nachteil? [Braucht zusätzlichen Speicher]
- Wieviel zusätzlicher Speicher? [Wussten wir beide nicht so genau, aber im Worst Case wohl etwa ein Verweis für jeden Eintrag (MUR) im Baum]