Diplomprüfung Praktische Informatik 14.06.2007

- Einführung in Datenbanken
- Data Mining
- Betriebssysteme

Einführung in Datenbanken

Seidl: Was ist denn eine Relation?
Ich: Eine Teilmenge eines kartesischen Produkts.
Seidl: Was braucht man denn noch so alles für eine Relation?
Ich: Name der Relation, Attributnamen und Wertbereiche (Domänen).
Seidl: Mehr.
Ich: Schlüssel. Ich kann Attribute als Schlüssel definieren.
Seidl: Was ist ein Schlüssel?
Ich: Eine Menge von Attributen, die alle anderen Attribute der Relation eindeutig bestimmen… vollständige funktionale Abhängigkeit = funktional abhängig und Schlüssel ist minimal.
Seidl: Was heißt minimal?
Ich: Linksreduziert.
Seidl: Angenommen ich habe eine Relation {[MatrNr, Name, VorlNr, Titel]}. Ist die gut?
Ich: Nein die ist nicht in 3NF. Da sind mehrere Konzepte drin realisiert. Wenn man schlecht modelliert hat wie hier, nimmt man Synthesealgorithmus.
Seidl: Mal kurz zeigen.
Ich: Die 3 FDs hingeschrieben:
fd1: {MatNr} → {Name}
fd2: {VorlNr} → {Titel}
fd3: {MatrNr, VorlNr} → {Name, Titel}
Kanonische Überdeckung machen, zuerst linksreduzieren, ist hier schon der Fall. Also noch rechtsreduzieren. Ich mache das mal am Beispiel „Titel“ in der fd3.
fd3’: {MatrNr, VorlNr} → {Name}
Titel $\in$ AttrHülle (F - fd3 $\cup$ fd3’, {MatrNr, VorlNr}). Damit fällt Titel raus. Name ebenfalls. Damit kann man fd3 komplett eliminieren. Letzter Schritt Zusammenfassen erübrigt sich. (Anmerkung: Um die kanonische Überdeckung zu lernen, sollte man mal eine Aufgabe aus Kemper Übungsbuch gemacht6 haben).
Seidl: Was mache ich mit den verbleibenden FDs?
Ich: Relationen draus bilden. Ich muss aber darauf achten, dass die Zerlegung verlustfrei und abhängigkeitserhaltend ist. Deswegen muss ich noch eine Relation hinzufügen, die die beiden verbindet, und die bilde ich aus dem ursprünglichen Schlüssel.
Seidl: Was bedeutet denn verlustfrei?
Ich: Wenn ich einen Join mache, muss ich die alte Relation rausbekommen. Eben deswegen muss ich noch die Relation {[MatrNr, VorlNr]} hinzufügen.
Seidl: Warum will man denn 3NF haben? Was ist so gut daran? Oder was ist an dem anderen Schema schlecht?
Ich: Können Anomalien auftreten. Einfüge-, Lösch- und Updateanomalie (jeweils erklärt). Bei Update-Anomalie hat man, selbst wenn man die Berücksichtigung aller Tupel erreichen könnte, immer noch unnötigen Speicherverbrauch.
Seidl: Aber wir haben doch nach dem Synthesealgorithmus viel mehr Relationen und z.B. MatrNr tritt jetzt in zwei Relationen auf. Ist das nicht auch erhöhter Speicherverbrauch?

Ich: Hm ja man hat tatsächlich einen Overhead dadurch. Aber entscheidend ist hier die Anzahl der Tupel. Die ist bei einzelnen Tabellen wesentlich geringer.
Seidl: Und wie sieht es mit dem Aufwand aus?
Ich: Durch die Segmentierung muss man Joins bilden und Joins sind teuer. Das kauft man sich leider damit ein.
Seidl: Das war jetzt der Entwurf. Wie komm ich denn an die Daten ran?
Ich: Mit Anfragesprachen, wie z.B. SQL. Ein SQL Statement sieht so aus: SELECT FROM WHERE
Seidl: Und was gibt's da noch für Ausdrücke?
Ich: z.B. SELECT DISTINCT für die Duplikateliminierung oder GROUP BY (Mist das hätte ich besser nicht gesagt, von der Gruppierung kommt man nämlich automatisch zu OLAP Anfragen… siehe später)
Seidl: Machen Sie mal eine Anfrage mit GROUP BY für Stud:{[MatrNr, Name, Sem, Hauptfach]}.
Ich: Ich nehme jetzt mal z.B. die Gruppierung nach Hauptfach. Hier muss man aufpassen, dass man die Attribute in GROUP BY auch in SELECT schreibt. (Schreib… Leider hakte es trotzdem bei der SELECT Anfrage, was muss da genau wo hin?)
Seidl: (hilft) Wie sieht denn die Ausgabe von dieser Anfrage aus?
Ich: Ich will so viele Zeilen wie Hauptfächer mit Anzahl der Studenten, die es belegen. Ahja, da oben muss count(MatrNr) rein. Also
SELECT Hauptfach, count(MatrNr) FROM Stud GROUP BY Hauptfach.
Seidl: Wie heißen solche Dinge wie count, sum, avg?
Ich: Aggregatfunktionen.

Data Mining

Seidl: Solche Anfragen hatten wir ja auch in der Vorlesung Data Mining bei OLAP. Welche Operationen gibt es denn da?
Ich: Drill Down, Roll up, Slice und Dice.
Seidl: Was ist Slice und Dice?
Ich: Stammel mir was zurecht und habe schon eine böse Vorahnung.
Seidl: Alles in SQL machen mit obiger Relation Stud.
Ich: (Genau das habe ich kommen sehen. Habe das zum Glück noch kurz vor der Prüfung angeschaut, aber angenehm war mir das nicht. In den Folien stehen die SQL Anfragen nicht. Im Kemper findet man aber was dazu.)
Roll Up und Drill Down macht man auch mit GROUP BY.
Seidl: Machen Sie mal.
Ich: (mehr oder weniger Attribute in GROUP BY und SELECT rein geschrieben).
Seidl: Und jetzt Slice.
Ich: Ich kann mich gar nicht an solche Folien erinnern. (Später sagte er mir, das sei eine Transferaufgabe gewesen.)
Seidl: (hilft) Was machen Sie denn beim Slicen, und wo würden Sie das am ehesten in die Anfrage einbauen?
Ich: Ja dann im WHERE-Teil (das musste ich dann zum Glück nicht mehr machen).
Seidl: OK. Was ist KDD?
Ich: Bla bla, Data Mining der wichtigste Schritt, dann Evaluierung der gefundenen Muster, Nutzbarmachung, Visualisierung (er hat mich ne ganze Zeit reden lassen).
Seidl: Was ist denn der Unterschied zwischen Klassifikation und Clustering?
Ich: Bla bla jeweils grob die Methoden und Ansätze für jedes Kapitel dargestellt.
Seidl: Bei den partitionierenden Verfahren hatten wir ja K-means und K-medoid. Was ist da der Unterschied?

Ich: Bei K-means braucht man Mittelwerte, funktioniert also nur mit kontinuierlichen Werten. Bei K-medoid sind die Repräsentanten Elemente aus der Menge selbst. Dafür führt er vollständige Suche durch mit $O(n^2)$ pro Element. Es gibt Verbesserungen: PAM, CLARANS (er hat mich auch hier lange reden lassen).

Seidl: Was ist denn hierarchisches Clustering?

Ich: Braucht man bei Clustern unterschiedlicher Dichte. Entweder ineinander enthalten, können aber auch nebeneinander sein. (Dendrogramme, Agglomeratives Clustering erklärt: Distanzfunktionen nicht nur zwischen einzelnen Punkten sondern auch zwischen Mengen). Dann gibt es noch OPTICS. (beim Erklären musste ich noch zu DBSCAN ausholen).

Seidl: Nun teilen sich DBSCAN und OPTICS ja einige Definitionen. Wie unterscheiden sich die beiden denn in ihrer Ausgabe?

Ich: Bei OPTICS kommt ein Erreichbarkeitsplot raus. Die Punkte sind ja geordnet, und da kann man die Cluster erkennen, und in diesem Cluster liegen noch dichtere Cluster drin (mal so ein Plot auf). DBSCAN macht ja gar keine visuelle Ausgabe sondern gibt eher nur Mengen aus.

Seidl: Ja OPTICS kann man ja auch als Vorstufe für DBSCAN nehmen.

Ich: Ich sehe das eher als Erweiterung.

Seidl: Dann frag ich mal so: Wo ist denn das ε im Plot?

Ich: Auf der senkrechten Achse. Ich kann das Diagramm quasi abschneiden (male waagerechte Linie ein, da funkt es). Achso wenn ich das ε zu klein wähle, sehe ich nur die sehr dichten Cluster. Wenn ich es zu hoch wähle, nur die weniger dichten Cluster.

Seidl: Wie sollte man ein gutes ε dann aus dem OPTICS-Plot wählen für DBSCAN?

Ich: So, dass ich alle Cluster sehen kann, also auf Höhe der Berge um die dichten Cluster.

Seidl: OK. Noch zum Subspace Clustering. Was ist denn da das Problem?

Ich: Einige Cluster sind nur in Subspaces sinnvoll zu erkennen. Wir hatten da ein Beispiel… Jetzt kann man aber nicht alle 2^d Subspaces untersuchen. Stattdessen macht man sich die Monotonieeigenschaft zu Nutze, ähnlich wie bei Apriori. Wenn man eine dichte Region hat, dann muss auch im Subspace eine dichte Region sein. Damit rechtfertigt man ein Bottom-Up-Vorgehen, wo man mit eindimensionalen Subspaces anfängt. Man streicht alle nicht-dichten raus und bildet dann alle möglichen Kombinationen daraus. Bei diesen schaut man wiederum, welche nicht dicht sind, und streicht sie usw.

Betriebssysteme

Seidl: Was ist ein Prozess?

Ich: Programm in Ausführung, mit Allokation von Ressourcen wie Speicherzuteilung und Register.

Seidl: Muss das OS von den Prozessen wissen?

Ich: Ja. Die Prozesse werden in der Prozesstabelle verwaltet. Da steht alles Mögliche drin...

Seidl: Muss der Prozess vom OS wissen?

Ich: Auch das. Die Prozesse müssen Systemaufrufe machen können.

Seidl: Was ist z.B. ein Systemaufruf?

Ich: Wenn ein Prozess eine Eingabe oder Ausgabe machen will auf einem Peripheriegerät, auf Dateien zugreifen, Speichervergrößerung anfordern,…

Seidl: Bei dem Zugriff auf Ressourcen kann es zu Deadlocks kommen. Was ist das?

Ich: Wenn sich Prozesse gegenseitig durch Halten von Ressourcen blockieren und keiner gibt was frei.

Seidl: Welche Bedingungen für Deadlock?

Ich: Die vier Bedingungen erklärt.

Seidl: Kann beim Speicher denn ein Deadlock auftreten?
Ich: Nein, da gibt es das Paging, der Speicher ist also entziehbar.
Seidl: Wie geht man in der Realität mit Deadlocks um?
Ich: (Anmerkung: hier sollte man NICHT einfach nur Vogelstrauß sagen!) Kommt drauf an, bei PC mit Windows oder Unix ignoriert man sie einfach, bei Sicherheitssystemen muss allerdings mit allen Mitteln versucht werden, Deadlocks zu vermeiden. Man kann das durch vorsichtige Vergabe von Ressourcen erreichen… bla bla sichere Zustände…
Seidl: Aber auch bei normalen Betriebssystemen hat man Bestrebungen unternommen, Deadlocks zu vermeiden. Denken Sie z.B. an Drucker.
Ich: Ja da nimmt man Spooling. Der Druckerdämon hat das alleinige Zugriffsrecht auf den Drucker.
Seidl: Welche Bedingung wird damit angegriffen?
Ich: Die Mutual Exclusion.
Seidl: Nun gibt es ja auch Multimedia OS. Was muss man da beachten?
Ich: Echtzeitanforderungen. Es gibt Deadlines. Meist hat man periodische Prozesse. Dadurch kann man in gewisser Weise die Last besser Voraussagen als bei interaktiven Systemen. Es muss gewährleistet sein, dass alle Prozesse in ihrer Periode abgearbeitet werden können.
Seidl: Was bedeutet Echtzeit im Allgemeinen?
Ich: Harte oder weiche Anforderungen. Bei Multimedia weiche: ärgerlich aber nicht tödlich wenn der Film mal hakt, bei Sicherheitssystemen harte Anforderungen.
Seidl: Welche Scheduling-Algorithmen gibt es dafür?
Ich: EDF und RMS. Beide erklärt.
Seidl: (bei RMS) Wenn der Prozess A die anderen unterbricht, wann sind denn die anderen dann dran?
Ich: Danach.
Seidl: Und wenn A so lange arbeitet?
Ich: Solange A Zeit braucht darf er natürlich. Aber der kann ja nicht ewig machen. Es gibt ja noch die generelle Bedingung, dass die Summe aller Prozesslängen pro Periodenlänge <1 ist. Wenn A ewig braucht und für die anderen nichts übrig bleibt, ist diese Bedingung augenscheinlich nicht erfüllt.
Seidl: Bei Multimedia ist ja noch was wichtig, wenn Sie mal an das Multi denken.
Ich: Ja die Synchronisation. Man will ja nicht den Ton neben dem Bild herlaufen haben.
Seidl: OK das reicht. Gehen Sie mal bitte kurz raus.

Zur Prüfung habe ich nichts zu sagen, was nicht andere schon gesagt hätten. Die Prüfung ist sehr angenehm verlaufen.
Tanenbaum ist ein anerkannter Autor, Prentice Hall ist ein anerkannter Verlag. Umso verwunderlicher, dass man ihm eine eher schlechte als rechte Übersetzung spendiert hat, und das schon in der zweiten Auflage. Wie dem auch sei, wenn man wenig Ahnung von der Materie hat, ist es schön, wenn man immer wieder Anknüpfpunkte kriegt und die Dinge mehrmals erzählt bekommt. Zum Wiederholen des Stoffs fand ich den Tausendseiter jedoch irgendwie unpraktisch. Da muss man sich bei Bedarf was anderes einfallen lassen und eine eigene Zusammenfassung machen.
Kemper ist angenehm, und auch hin und wieder mal das Übungsbuch zur Hand nehmen. Man kann getrost auch die 5. Auflage statt der 6. kaufen, ich habe keinen Unterschied festgestellt. Die Vorlesungsfolien von Prof. Seidl kommen aus professioneller Hand und die deutsche Unterstützung in Form des Ester/Sander-Buches ist auch eine gute Wahl.

Ich danke allen Leuten, die ein Prüfungsprotokoll geschrieben haben!