

Beauty is our Business

Organisatorisches und Einführung

19.4.2007

Felix Naumann

Das Motto

2

... Wenn wir uns klarmachen, dass der Kampf gegen Chaos, Durcheinander, und unbeherrschte Kompliziertheit eine der größten Herausforderungen der Informatik ist, müssen wir zugestehen:

„Beauty is our Business“.

Edsger W. Dijkstra, 1978

Auch:

“Computer Science is no more about computers than astronomy is about telescopes.”

Motivation

3

- Vorbereitung auf das Master Studium
 - Wissenschaftlich Arbeiten
- Vorbereitung auf den Beruf
 - Ideen „verkaufen“
- Interesse an den Themen

Überblick

4

- ➔ ■ Vorstellung der Arbeitsgruppe
- Organisatorisches
- Ziele des Seminars
- Zeitlicher Ablauf
- Kurzvorstellung der Themen



5

Felix Naumann

- naumann@hpi.uni-potsdam.de
- Raum A-1.13 (über Frau Pamperin)
- HU, IBM Almaden, DFG/JP/HU, HPI

Universität Potsdam, Hasso-Plattner-Institut

- Fachgebiet „Informationssysteme“

Forschungsthemen (gleich mehr):

- Datenqualität
- Informationsintegration
- Peer Data Management
- Search
- <http://www.hpi.uni-potsdam.de/~naumann/>

Felix Naumann | SE Beauty is our Business | SS 07

6

Leitung

- Prof. Felix Naumann
- Patricia Hobro (Sekretariat)

Wissenschaftliche Mitarbeiter

- Jens Bleiholder (Datenfusion, HumMer, FuSem)
- Armin Roth (PDMS, System P)
- Melanie Weis (Duplikaterkennung, DogmatiX, XClean)
- Jana Bauckmann (Data Profiling, Aladin)
- Alexander Albrecht (PIM, IIS)
- Frank Kaufer (Matching, Forschungskolleg)

Tutoren

- Christoph Böhm, Karsten Draba, Dustin Lange, Matthias Weidlich

Felix Naumann | SE Beauty is our Business | SS 07

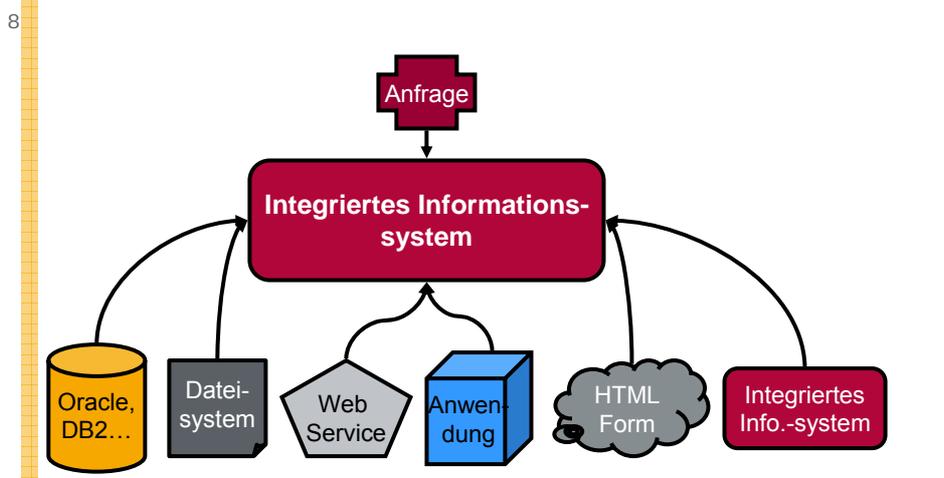
Was sind Informationssysteme?

7

```

<buch>
  <isbn>0-201-318051</isbn>
  <titel>XML und Datenbanken</titel>
  <autor>Klettke/Meyer</autor>
</buch>
  
```

Integrierte Informationssysteme



Schematische und Daten-Heterogenität

9

Variante 1

Männer	
Vorname	Nachname
Felix	Naumann
Jens	Bleiholder

Frauen	
Vorname	Nachname
Melanie	Weis
Jana	Bauckmann

Variante 2

Personen			
Vorname	Nachname	Männl.	Weibl.
Felix	Naumann	Ja	Nein
Jens	Bleiholder	Ja	Nein
Melanie	Weis	Nein	Ja
Jana	Bauckmann	Nein	Ja

Variante 3

Personen		
Vorname	Nachname	Geschlecht
Felix	Naumann	Männlich
Jens	Bleiholder	Männlich
Melanie	Weis	Weiblich
Jana	Bauckmann	Weiblich

Felix Naumann | SE Beauty is our Business | SS 07

Schematische und Daten-Heterogenität

10

Variante 1

Männer	
Vorname	Nachname
Felix	Naumann
Jens	Bleiholder

Frauen	
Vorname	Nachname
Melanie	Weis
Jana	Bauckmann

Variante 2

Personen			
FirstNa	Name	male	femal
Felix	Naumann	Ja	Nein
Jnes	Bleiho.	Ja	Nein
Melanie	Weiß	Nein	Ja
Jana	bauckman	Nein	Ja

Variante 3

Personen		
VN	NN	SEX
F.	Naumann	Männlich
J.	Bleiholder	Männlich
M.	Weis	Weiblich
J.	Bauckmann	Weiblich

Felix Naumann | SE Beauty is our Business | SS 07

11

Variante 1

Heterogenität	
Personen	Produkte
Person A	Produkt X
Person B	Produkt Y

Heterogenität	
Personen	Produkte
Person C	Produkt Z
Person D	Produkt W

Variante 2

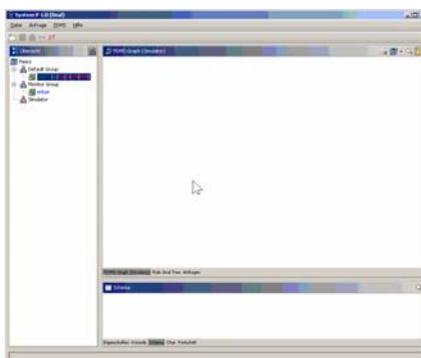
Heterogenität			
Personen	Produkte	Orte	Zeiten
Person A	Produkt X	Ort 1	Zeit 1
Person B	Produkt Y	Ort 2	Zeit 2
Person C	Produkt Z	Ort 3	Zeit 3
Person D	Produkt W	Ort 4	Zeit 4

Variante 3

Heterogenität		
Personen	Produkte	Orte
Person A	Produkt X	Ort 1
Person B	Produkt Y	Ort 2
Person C	Produkt Z	Ort 3
Person D	Produkt W	Ort 4

12

- Informationsintegration
 - Schema Matching
 - Duplikaterkennung
 - Datenfusion
- Datenqualität
- Peer Data Management
- Life Sciences: Aladin
- Search



Lehrveranstaltungen

13

Vorlesungen

- DBS I
- DBS II
- Informationsintegration
- ...

Seminare

- Beauty is our Business
- Datenreinigung
- ...



Extending the Database Relational Model to Capture More Meaning

E. F. COGG

IBM Research Laboratory

During the last three or four years several investigations have been exploring "semantic models" for relational databases. The intent is to capture in a form or two formalized form of the meaning of the data so that database design can become more systematic and the database system itself can be more intelligent. The main driver are clear:

- (1) the search for meaningful data that are useful to people, other computers, systems;
- (2) the search for meaningful data that are useful to people, other computers, systems.

In the past we have attempted to do this by using certain codes and symbols marking data in the structure of the data for insertion, deletion, and retrieval as well as new algebraic systems.

Key words and phrases: relation, relational database, relational model, relational algebra, database, data model, database system, data structure, database model, knowledge representation, knowledge base, conceptual model, conceptual schema, entity model.

DB Concepts: 3.10, 3.11, 3.12, 4.10, 4.11, 4.12, 4.13

1. INTRODUCTION

The relational model for hierarchical databases [1] was conceived ten years ago, primarily as a tool to free users from the restrictions of having to deal with the complexity of changing implementation details. This implementation independence was achieved with the power of the algebraic operations on a very primitive and the open between relations have stimulated research in hierarchical, multi-level, and other models. The relational model has also provided an architectural design for the design of database and some general purpose database management systems such as MACADAM [2], PILEY [3], RDM/SCM [4], MACRODM [5], INGRES [6], QRE [4], and others [7].

During the last few years intensive investigations have been aimed at capturing

Permission to copy without fee all or part of this material is granted provided that the copies are not made for advertising or promotional purposes, for creating new collective works, or for resale. Copyright © 1979 by IBM Research Laboratory, 5600 Cedar Road, San Jose, CA 95128. 01979-0198 IBM Research Laboratory, 5600 Cedar Road, San Jose, CA 95128.

Überblick

14

- Vorstellung der Arbeitsgruppe
- Organisatorisches
- Ziele des Seminars
- Zeitlicher Ablauf
- Kurzvorstellung der Themen



Anmeldung / Teilnahme

15

Beauty is our business

Beschreibung

In diesem Seminar lernen Sie wissenschaftliche Arbeiten kritisch aber mit Genuß zu lesen, in einem Vortrag verständlich aber auch vergnüglich wiederzugeben und in einer Ausarbeitung ansprechend zu formulieren.

„Wenn wir uns klarmachen, daß der Kampf gegen Chaos, Durcheinander und unbeherrschte Kompliziertheit eine der größten Herausforderungen der Informatik ist, müssen wir zugestehen: Beauty is our Business.“ Edsger W. Dijkstra, 1978

Das Seminar findet in intimen Kreis statt: Es werden nur sechs Bachelorstudenten zugelassen. Die Anmeldung erfolgt per kurzer E-Mail direkt an mich. Ich werde alle zum ersten Termin einladen, eine Nachruckerliste führen, und danach wird festgelegt, wer teilnimmt.

Per E-Mail angemeldet haben sich

- Philipp Dobrigkeit
- Stefan Krumnow
- Daniel Hefenbrock
- Matthias Pohl
- Tobias Flach

- Durchschnitt DBS I: 1,3 (2,6)

Anmeldungsprocedere bei vielen Teilnehmern

- Nach heutiger Vorstellung: Bis Montag **verbindliche** Zusage per E-Mail an mich
 - Mit Wunschthema und Zweit- und Drittwunsch
- Bei zu vielen Teilnehmern
 - Verlosung von Plätzen
 - Vorrang der fünf bereits angemeldeten

Felix Naumann | SE Beauty is our Business | SS 07

Seminarleistungen

16

Lesen

- Paper lesen und verstehen
- Verwandte Literatur lesen und verstehen
- Mindestens eine individuelle Besprechung mit mir
- Kurzvorstellung der Literatur am 10.5.

Vortragen

- Mindestens eine Folien-Besprechung mit mir
 - Spätestens 1 Woche vor Vortrag
- 30 min. Vortrag am jeweiligen Termin
 - + 15 min Diskussion
- Aktive Teilnahme an anderen Vorträgen

Ausarbeitung

- Mindestens eine Gliederungs-Besprechung mit mir
 - Spätestens 2 Wochen vor Abgabetermin
 - Max. 10-seitige Ausarbeitung bis zum 30.7.2007
 - Unter Verwendung der LaTeX-Vorlage im WWW

Aktive Teilnahme an sämtlichen gemeinsamen Terminen

Felix Naumann | SE Beauty is our Business | SS 07

Feedback

17

Fragen bitte jederzeit!

- Während des Seminars
- Während individueller Besprechungen
 - Termin bei Pat Hobro ausmachen
 - brigitte.hobro@hpi.uni-potsdam.de
 - 0331 / 5509 280
 - Muss nicht zur Sprechstunde sein!
- Sprechstunde
 - Dienstags 15:00 – 16:00
 - Raum A-1.13
 - Am liebsten mit Anmeldung
- Email: naumann@hpi.uni-potsdam.de

Felix Naumann | SE Beauty is our Business | SS 07

Überblick

18

- Vorstellung der Arbeitsgruppe
- Organisatorisches
- Ziele des Seminars
- Zeitlicher Ablauf
- Kurzvorstellung der Themen



Felix Naumann | SE Beauty is our Business | SS 07

Wissenschaftliche Texte lesen

19

- Fachartikel
 - Journale, Konferenzen und Workshops
 - Entstehung:
 - Forschungsvorhaben
 - Begutachtungsprozess
- Struktur eines Artikels
 - Kritisches Lesen
 - Experimente
- Literatur
 - (Online-) Recherche
- Englisch

Felix Naumann | SE Beauty is our Business | SS 07

Wissenschaftlichen Vortrag halten

20

- Gliederung
 - Die Kunst des Weglassens
- Foliengestaltung
 - Powerpoint
 - Overhead
- Zeit einhalten
- Techniken zur Vorbereitung
- Techniken während des Vortrags

Felix Naumann | SE Beauty is our Business | SS 07

Ausarbeitung schreiben

21

- Gliederung
- Schreibstil
 - Objektives
 - Subjektives
- Plagiate
- LaTeX

Überblick

22

- Vorstellung der Arbeitsgruppe
- Organisatorisches
- Ziele des Seminars
- ➔ ■ Zeitlicher Ablauf
- Kurzvorstellung der Themen



Termine und Themenvergabe		Hasso Plattner Institut	
Termin	Thema	Vortrag	Folien
23 19.4.2007	Einführung	Felix Naumann	pdf
26.4.2007	Wissenschaftliche Texte Lesen	Felix Naumann	pdf
10.5.2007	Literaturkritik / Diskussion	Alle	
17.5.2007	Vortragstechniken	Felix Naumann	pdf
7.6.2007	Vortrag 1: Mariposa (pdf) Vortrag 2: Fagins Algorithmus (pdf)		
14.6.2007	Einführung in LaTeX	Felix Naumann	pdf
21.6.2007	Vortrag 3: Enough Already in SQL (pdf) Vortrag 4: Sorted Neighborhood (pdf)		
28.6.2007	Vorstellung der Gliederungen & Tipps zur Ausarbeitung	Alle	
5.7.2007	Vortrag 5: Source Capabilities (pdf) Vortrag 6: Data Mining (pdf)		
30.7.2007	Abgabe der Ausarbeitungen	Alle	

Felix Naumann | SE Beauty is our Business | SS 07

24



Überblick

- Vorstellung der Arbeitsgruppe
- Organisatorisches
- Ziele des Seminars
- Zeitlicher Ablauf
- ➔ ■ Kurzvorstellung der Themen



Felix Naumann | SE Beauty is our Business | SS 07

Allgemeines

25

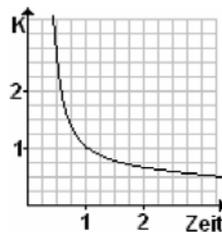
- Themen aus dem Umfeld
 - Datenbanken (DB)
 - Informationsintegration (II)
 - Business Intelligence (BI)
- Sehr gute, oft wegweisende paper
 - Jeweils mindestens eine wirklich gute Idee
 - Gut geschrieben

Felix Naumann | SE Beauty is our Business | SS 07

Mariposa

26

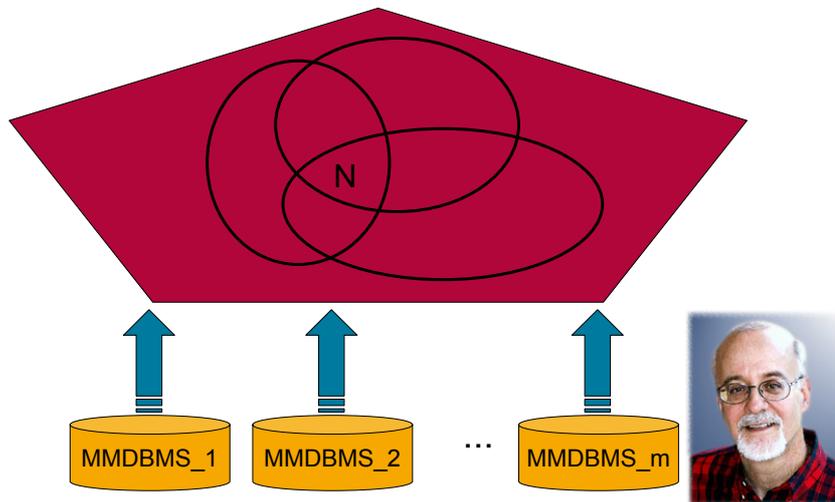
- Neuartige Architektur für ein weit verteiltes Informationssystem
- Mikroökonomisches Prinzip zur Anfrageoptimierung
 - Datenhaltung
 - Anfragebearbeitung



Felix Naumann | SE Beauty is our Business | SS 07

Fagins Algorithmus

27



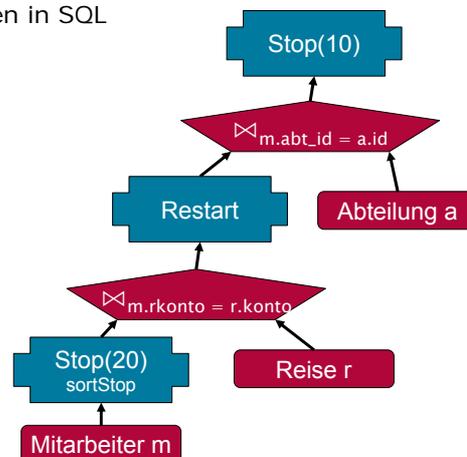
Felix Naumann | SE Beauty is our Business | SS 07

„Enough Already“ in SQL

28

■ First-N und Top-N Techniken in SQL

- Syntax & Semantik
- Neue Operatoren
- Optimierung
- Evaluation

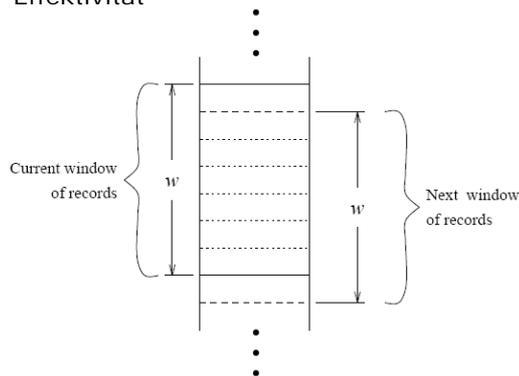


Felix Naumann | SE Beauty is our Business | SS 07

Sorted Neighborhood

29

- Datenreinigung und Duplikaterkennung
 - Effizienz
 - Effektivität



Felix Naumann | SE Beauty is our Business | SS 07

Source Capabilities

30

- Gebundene und freie Variablen

$v_1(\text{Song}, \text{CD})$

<Friends, Life>

<Friends, Love>

$v_2(\text{CD}, \text{Artist}, \text{Price})$

<Love, Lucy, \$15>

<Story, Snoopy, \$14>

$v_3(\text{CD}, \text{Artist}, \text{Price})$

<Story, Lucy, \$13>

<Love, Snoopy, \$10>

<Life, Charlie, \$8>

Bastelaufgabe 1:
Wie teuer ist die billigste CD mit einem Song namens "Friends"?

Bastelaufgabe 2:
Welches ist die billigste CD mit einem Song namens "Friends", die Sie anfragen können?



Felix Naumann | SE Beauty is our Business | SS 07

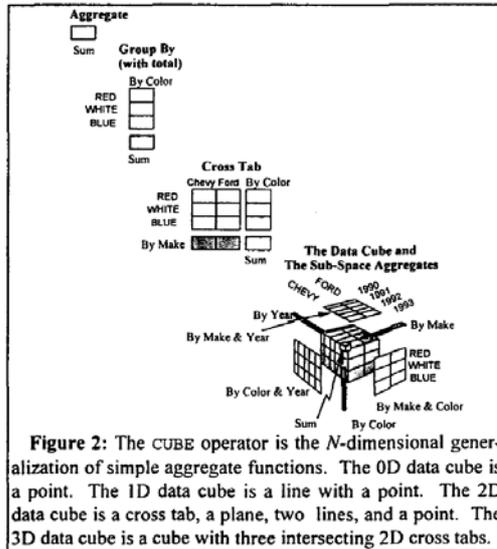
Data Cube

31

- Data Warehouses
- Effiziente Gruppierung und Aggregation



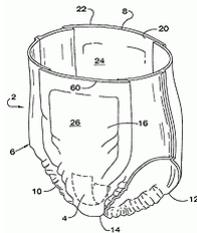
Felix Naumann | SE Beauty is our Business | SS 07



Data Mining

32

- Bahnbrechendes Papier
- Setzte intensive Forschung zu Data Mining in Gang
- Zwei Algorithmen zur schnellen Entdeckung von Assoziationsregeln
 - Apriori
 - AprioriTid



Felix Naumann | SE Beauty is our Business | SS 07