Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

Beauty is our Business

Organisatorisches und Einführung

19.4.2007

Felix Naumann

Das Motto

... 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

Forschungsgruppe Informationsintegration

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/

Unsere Arbeitsgruppe

6

Leitung

- Prof. Felix Naumann
- Patricia Hobro (Sekretariat)

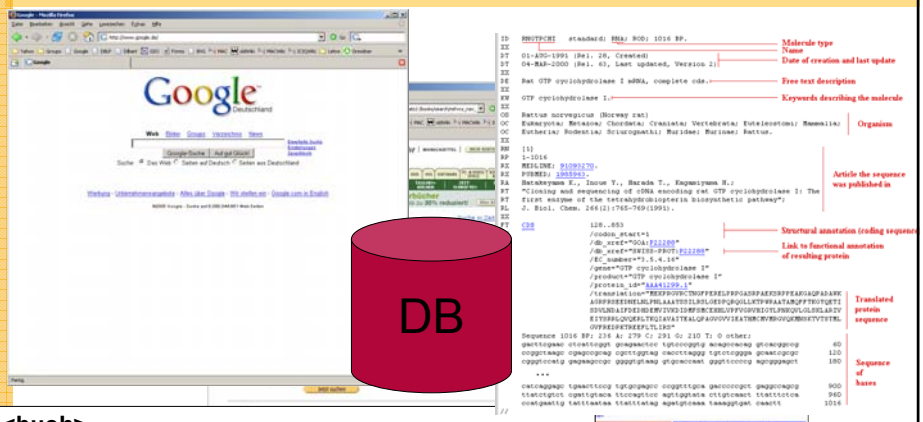
Wissenschaftliche Mitarbeiter

- 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

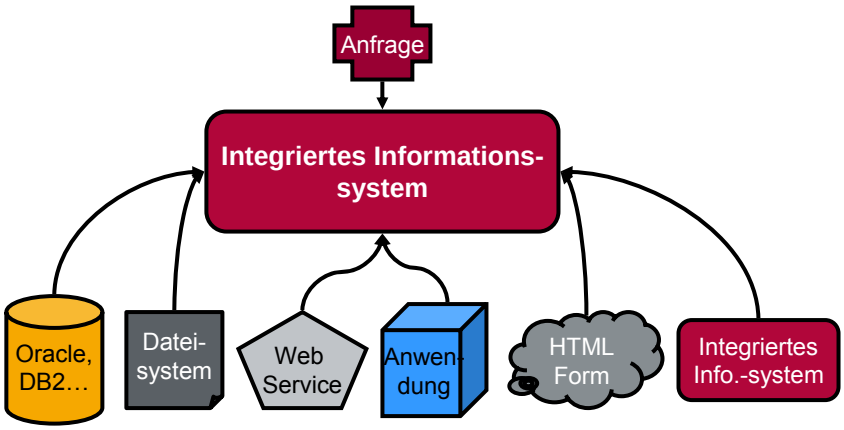
- Christoph Böhm, Karsten Draba, Dustin Lange, Matthias Weidlich

Was sind Informationssysteme?



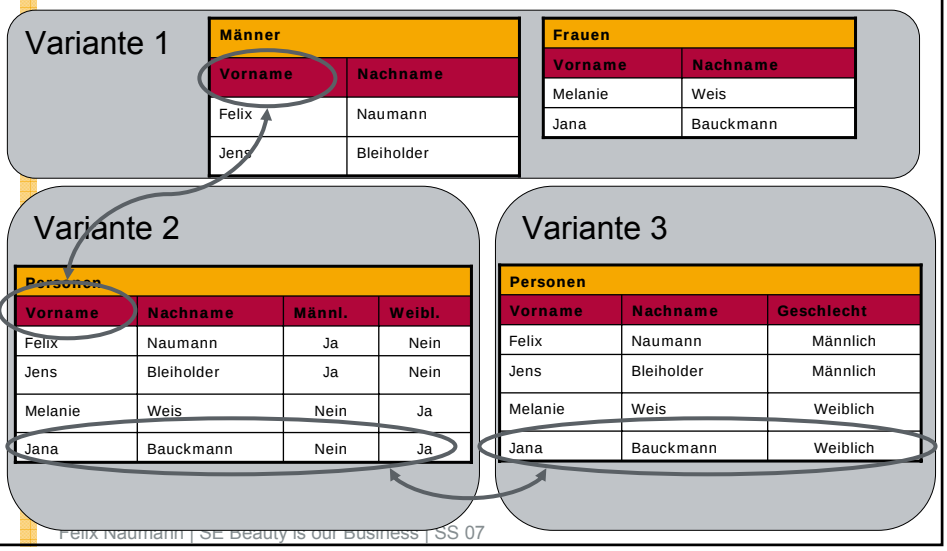

```
<buch>
        <isbn>0-201-318051</isbn>
        <titel>XML und Datenbanken</titel>
        <autor>Klettke/Meyer</autor>
</buch>
```


Integrierte Informationssysteme



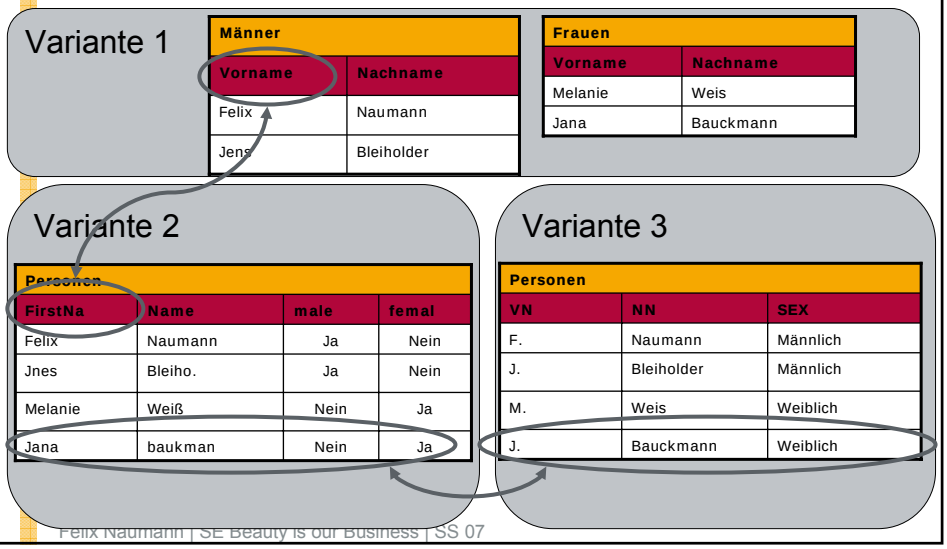
Schematische und Daten-Heterogenität

9



Schematische und Daten-Heterogenität

10



Schematische und Daten-Heterogenität

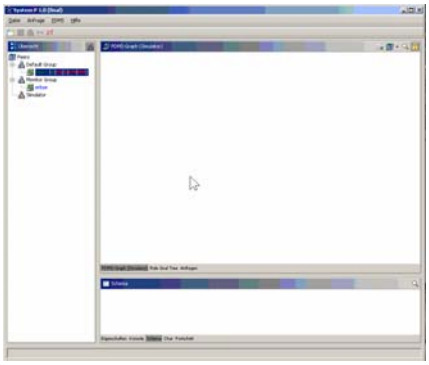
Variante 1

Variante 2

Variante 3

Weitere Forschungsthemen

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



Lehrveranstaltungen

13


Extending the Database Relational Model to Capture More Meaning

E. F. CODD

IBM Research Laboratory

Vorlesungen

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

Seminare

- Beauty is our Business
- Datenreinigung
- ...

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 Nachrückerliste 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

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

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

Überblick

18

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

Wissenschaftliche Texte lesen

19

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

Wissenschaftlichen Vortrag halten

20

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

Ausarbeitung schreiben

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

Überblick

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

23

Termine und Themenvergabe

Termin	Thema	Vortrag	Folien
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	

Hasso Plattner Institut

24

Überblick

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

Allgemeines

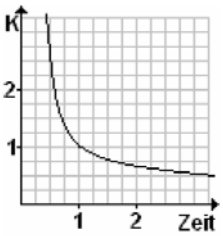
25

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

Mariposa

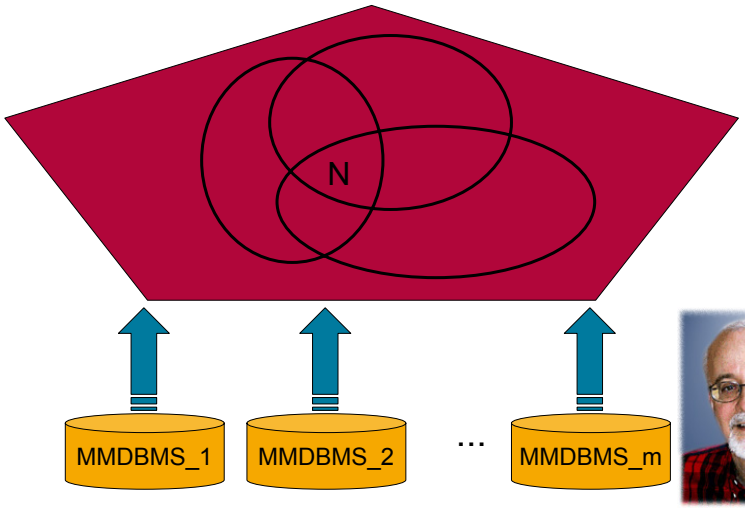
26

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



Fagins Algorithmus

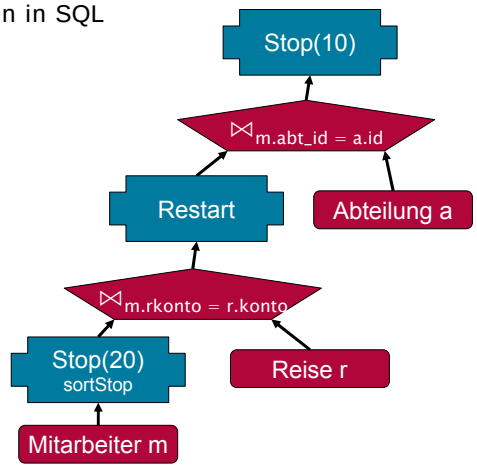
27



„Enough Already“ in SQL

28

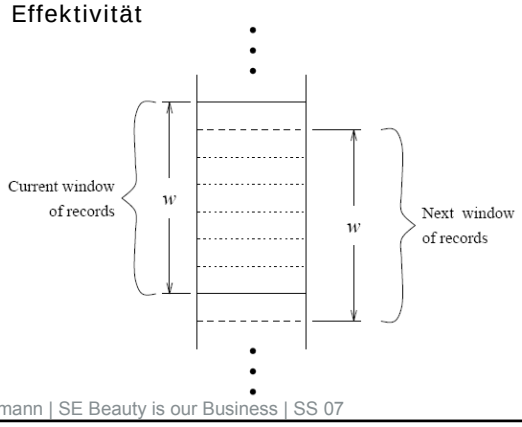
- First-N und Top-N Techniken in SQL
  - Syntax & Semantik
  - Neue Operatoren
  - Optimierung
  - Evaluation



Sorted Neighborhood

29

- Datenreinigung und Duplikaterkennung
  - Effizienz
  - Effektivität



Source Capabilities

30

- Gebundene und freie Variablen

v_1(Song, CD)

<Friends, Life>

<Friends, Love>

v_2(CD, Artist, Price)

<Love, Lucy, $15>

<Story, Snoopy, $14>

v_3(CD, Artist, 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?

Data Cube

- Data Warehouses
- Effiziente Gruppierung und Aggregation

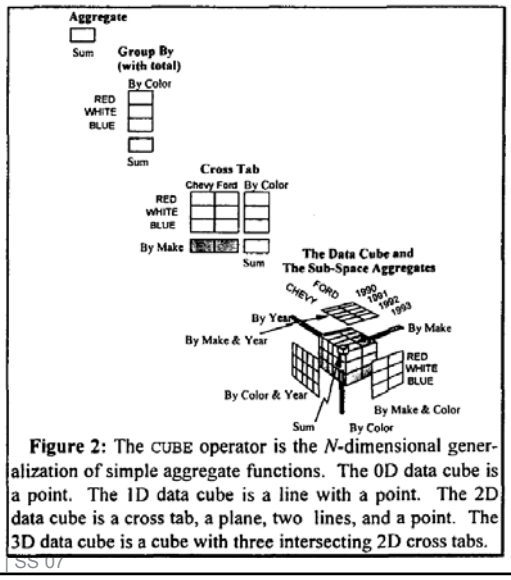


Figure 2: The CUBE operator is the *N*-dimensional generalization of simple aggregate functions. The 0D data cube is a point. The 1D data cube is a line with a point. The 2D data cube is a cross tab, a plane, two lines, and a point. The 3D data cube is a cube with three intersecting 2D cross tabs.

Data Mining

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