Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

Beauty is our Business

Organisatorisches und Einführung

22.10.2007

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

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

Überblick

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

Forschungsgruppe Informationsintegration

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

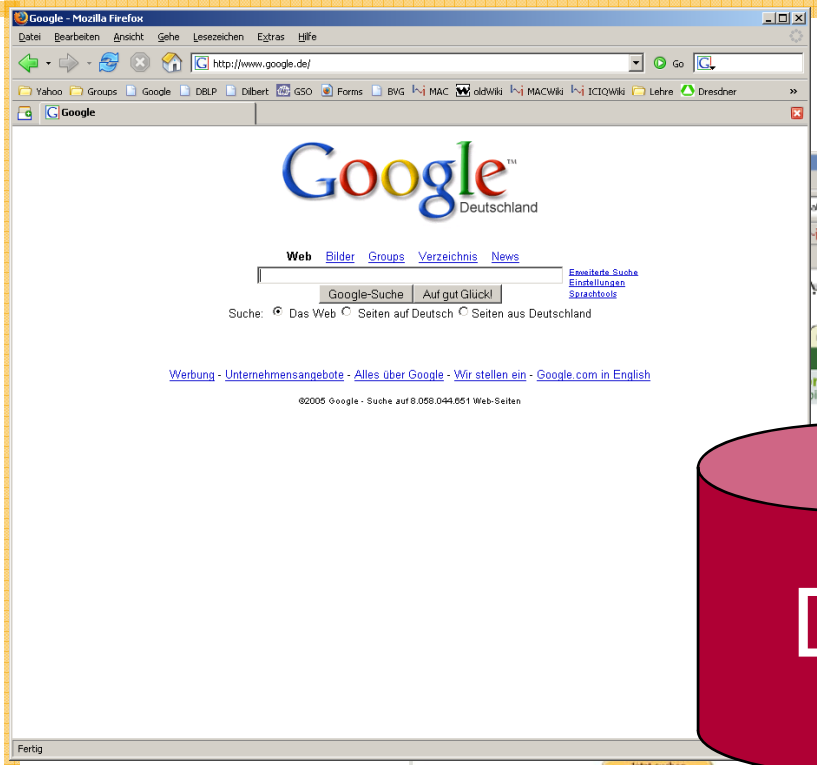
- Wissenschaftliche Mitarbeiter
  - Alexander Albrecht: ETL, Schema Mapping, PIM
  - Jana Bauckmann: Data Profiling, Aladin
  - Jens Bleiholder: Data Fusion, HumMer & FuSem
  - Paul Führing: DQ Assessment, Viqtor
  - Frank Kaufer: Schema and Ontology Matching
  - Armin Roth: Peer-Data-Management, System P
  - Melanie Weis: Duplicate Detection
- Student assistants
  - Karsten Draba: HumMer & FuSem
  - N.N: Ranking, SPRINT
  - Tobias Flach: Aladin Project
  - N.N.: System P

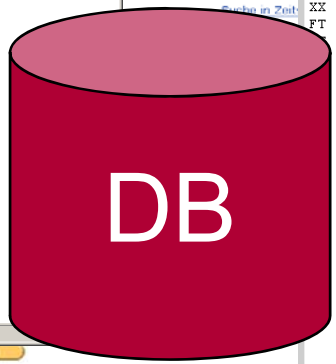






Was sind Informationssysteme?



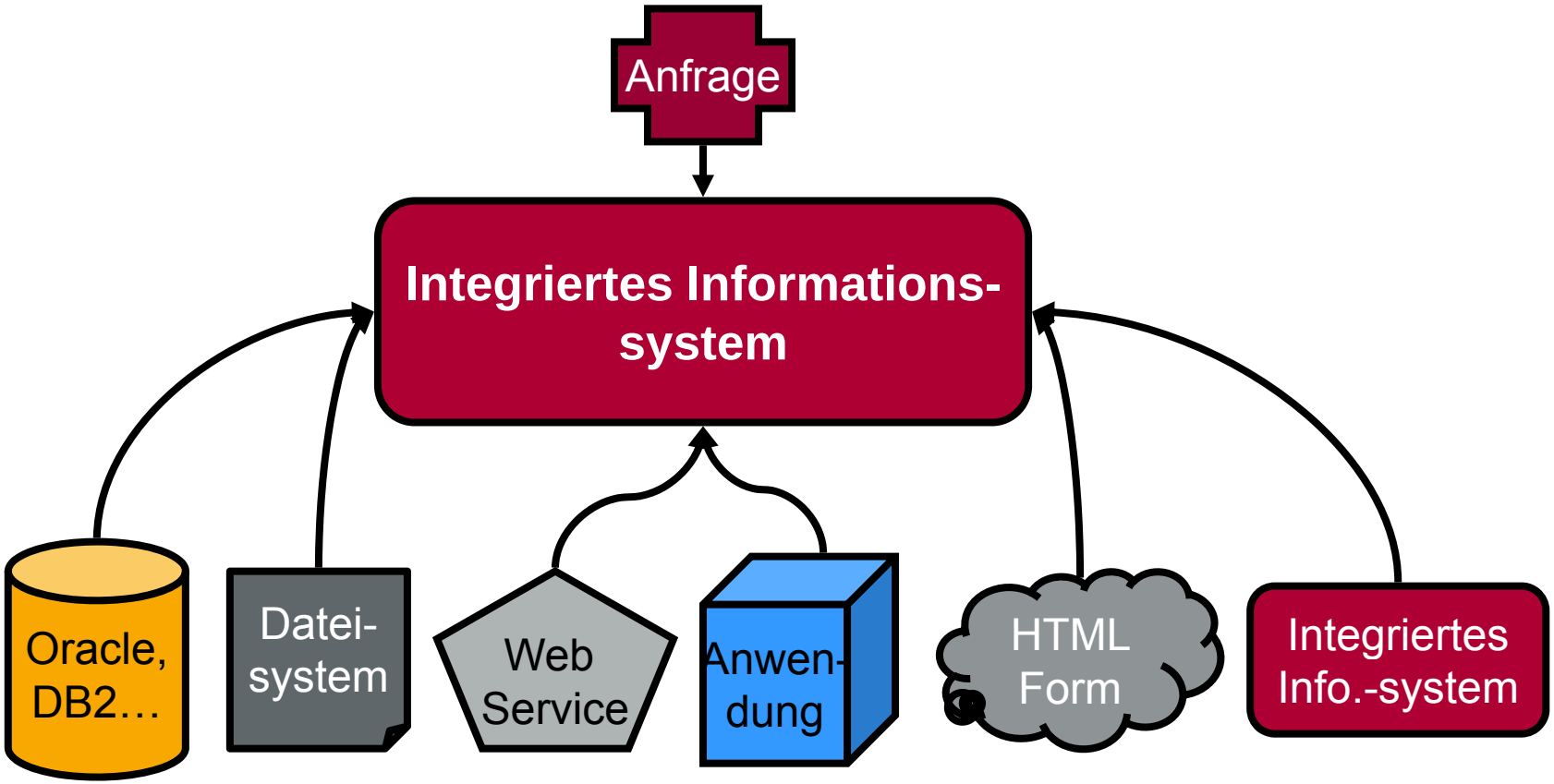



```
ID   RNGTPCHI   standard; RNA; ROD; 1016 BP.
XX
DT   01-AUG-1991 (Rel. 28, Created)
DT   04-MAR-2000 (Rel. 63, Last updated, Version 2)
XX
DE   Rat GTP cyclohydrolase I mRNA, complete cds.
XX
KW   GTP cyclohydrolase I.
XX
OS   Rattus norvegicus (Norway rat)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Rattus.
XX
RN   [1]
RP   1-1016
RX   MEDLINE; 91093270.
RX   PUBMED; 1985963.
RA   Hatakeyama K., Inoue Y., Harada T., Kagamiyama H.;
RT   "Cloning and sequencing of cDNA encoding rat GTP cyclohydrolase I: The
RT   first enzyme of the tetrahydrobiopterin biosynthetic pathway";
RL   J. Biol. Chem. 266(2):765-769(1991).
XX
FT   CDS             128..853
FT                   /codon_start=1
FT                   /db_xref="GOA:P22288"
FT                   /db_xref="SWISS-PROT:P22288"
FT                   /EC_number="3.5.4.16"
FT                   /gene="GTP cyclohydrolase I"
FT                   /product="GTP cyclohydrolase I"
FT                   /protein_id="AAA41299.1"
FT                   /translation="MEKPRGVRCTNGFPERELPRPGASRPAEKSRPPEAKGAQPADAWK
FT                   AGRPRSEEDNELNLPNLAAAYSSILRSLGEDPQRQGLLKTPWRAATAMQFFTKGYQETI
FT                   SDVLNDAIFDEDHDEMVIVKDIDMFSMCEHHLVPFVGRVHIGYLPNKQVLGLSKLARIV
FT                   EIYSRRLQVQERLTKQIAVAITEALQPAGVGVVIEATHMCMVMRGVQKMNSKTVTSTML
FT                   GVFREDPKTREEFLTLIRS"
Sequence 1016 BP; 236 A; 279 C; 291 G; 210 T; 0 other;
gacttcgaac ctcattcggt gcagaactcc tgtcccggtg acagccacag gtcacggccg        60
ccggctaagc cgagccgcag cgcttggtag caccttaggg tgtctcggga gcaatcgcgc       120
cgggtccatg gagaagccgc ggggtgtaag gtgcaccaat gggttccccg agcgggagct       180
   ...
catcaggagc tgaacttccg tgtgcgagcc ccggtttgca gacccccgct gaggccagcg       900
ttatctgtct cgattgtaca ttccagttcc agttggtata cttgtcaact ttatttctca       960
ccatgaattg tatttaataa ttatttatag agatgtcaaa taaaggtgat caacttt         1016
//
```

Molecule type
Name
Date of creation and last update
Free text description
Keywords describing the molecule
Organism
Article the sequence was published in
Structural annotation (coding sequence)
Link to functional annotation of resulting protein
Translated protein sequence
Sequence of bases


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

Integrierte Informationssysteme



Schematische und Daten-Heterogenität

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

Schematische und Daten-Heterogenität

Variante 1

Frauen	
Vorname	**Nachname**
Melanie	Weis
Jana	Bauckmann

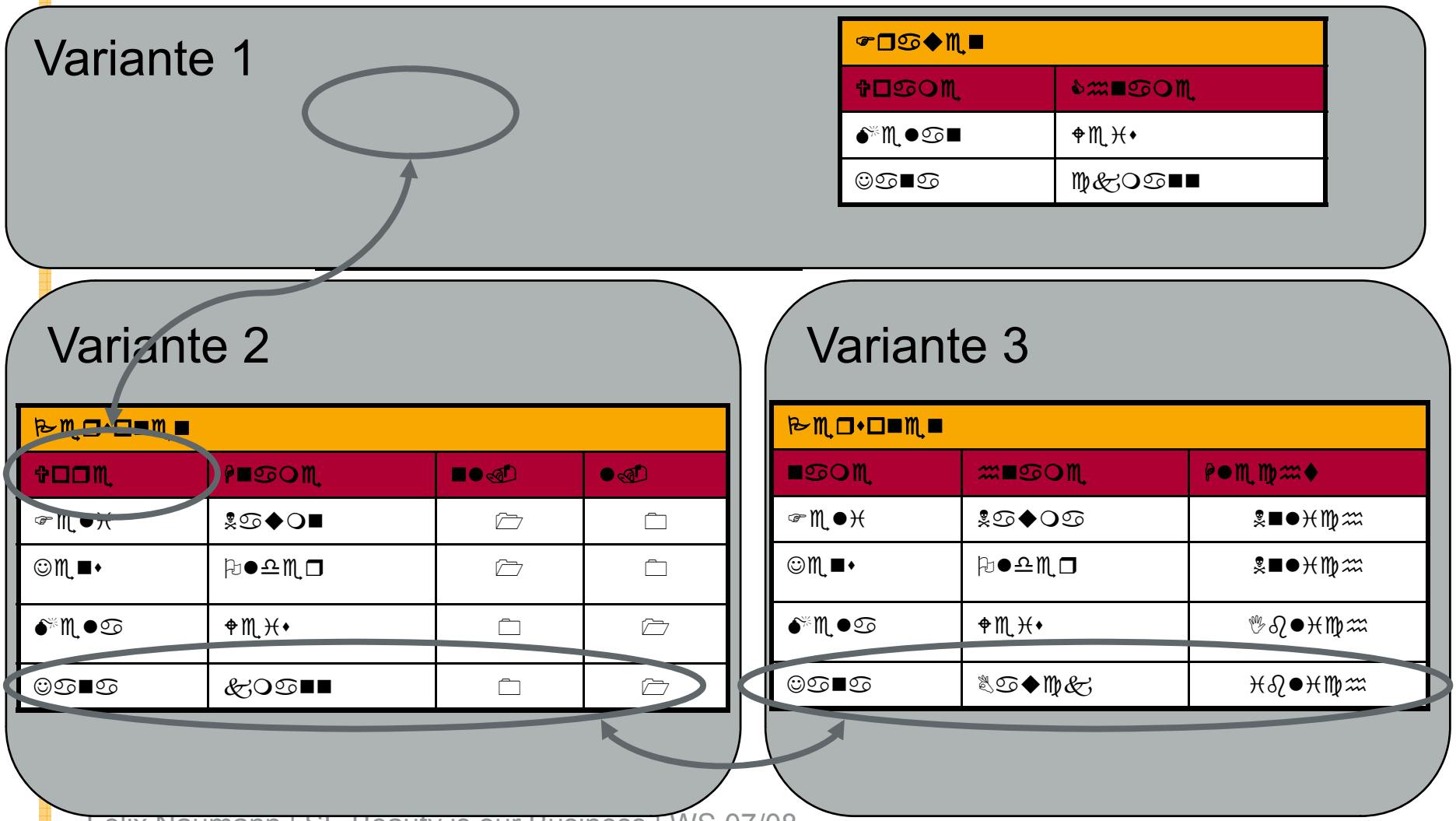
Variante 2

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

Variante 3

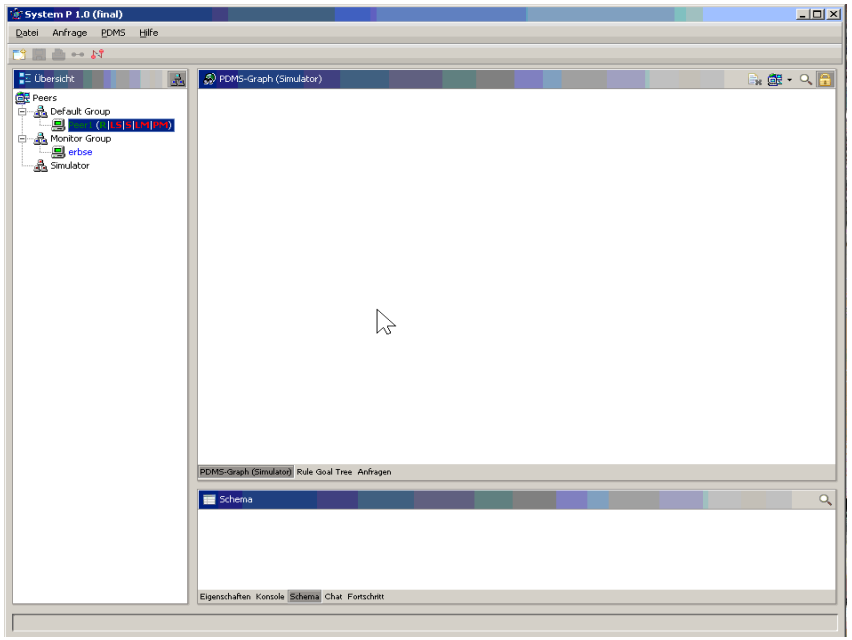
Personen		
VN	**NN**	**SEX**
F.	Naumann	Männlich
J.	Bleiholder	Männlich
M.	Weis	Weiblich
J.	Bauckmann	Weiblich

Schematische und Daten-Heterogenität



Weitere Forschungsthemen

- Informationsintegra
  - Schema Matchi
  - Duplikaterkennung
  - Datenfusion
- Datenqualität
- Peer Data Management
- Life Sciences: Aladin
- Search



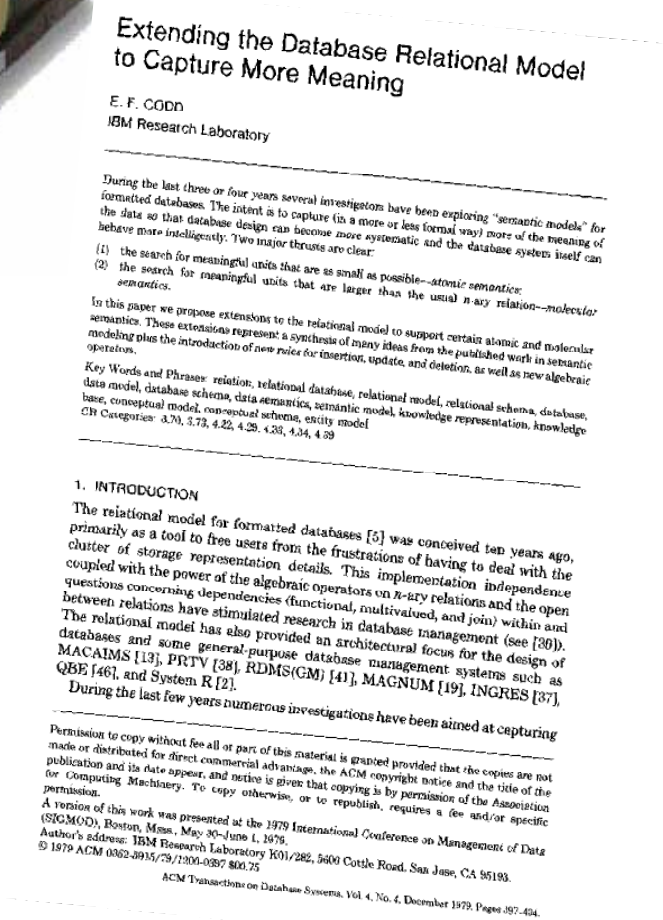
Lehrveranstaltungen

Vorlesungen

- DBS I
- DBS II
- Informationsintegrati
- …

Seminare

- Beauty is our Business
- Datenreinigung / Schema Mathing
- www.prominentpeople.info
- …



Extending the Database Relational Model to Capture More Meaning

E. F. CODD

IBM Research Laboratory

During the last three or four years several investigators have been exploring "semantic models" for formatted databases. The intent is to capture (in a more or less formal way) more of the meaning of the data so that database design can become more systematic and the database system itself can behave more intelligently. Two major thrusts are clear:

(1) the search for meaningful units that are as small as possible—*atomic semantics*;
(2) the search for meaningful units that are larger than the usual *n*-ary relation—*molecular semantics*.

In this paper we propose extensions to the relational model to support certain atomic and molecular semantics. These extensions represent a synthesis of many ideas from the published work in semantic modeling plus the introduction of *new rules* for insertion, update, and deletion, as well as new algebraic operators.

Key Words and Phrases: relation, relational database, relational model, relational schema, database, data model, database schema, data semantics, semantic model, knowledge representation, knowledge base, conceptual model, conceptual schema, entity model
CR Categories: 3.70, 3.73, 4.22, 4.29, 4.33, 4.34, 4.39

1. INTRODUCTION

The relational model for formatted databases [5] was conceived ten years ago, primarily as a tool to free users from the frustrations of having to deal with the clutter of storage representation details. This implementation independence coupled with the power of the algebraic operators on *n*-ary relations and the open questions concerning dependencies (functional, multivalued, and join) within and between relations have stimulated research in database management (see [30]). The relational model has also provided an architectural focus for the design of databases and some general-purpose database management systems such as MACAIMS [13], PRTV [38], RDMS(GM) [41], MAGNUM [19], INGRES [37], QBE [46], and System R [2].

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

A version of this work was presented at the 1979 International Conference on Management of Data (SIGMOD), Boston, Mass., May 30–June 1, 1979.
Author's address: IBM Research Laboratory K01/282, 5600 Cottle Road, San Jose, CA 95193.

Überblick

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

Vorstellung Teilnehmer

- Florian Reinhart
- Alexander Koglin
- Dandy Fenz
- Ziawasch Abedjan
- Arian Treffer
- Nico Rehwaldt

Seminarleistungen

Lesen
- Paper lesen und verstehen
- Verwandte Literatur lesen und verstehen
- Mindestens eine individuelle Besprechung mit mir
- Kurzvorstellung der Literatur (je 10min)

1/5 Note

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

2/5 Note

Ausarbeitung
- (Mindestens eine Gliederungs-Besprechung mit mir)
  - Spätestens 2 Wochen vor Abgabetermin
  - 8-10-seitige Ausarbeitung bis zum 29.02.2008
  - Unter Verwendung der LaTeX-Vorlage im WWW

2/5 Note

Aktive Teilnahme an sämtlichen gemeinsamen Terminen

Feedback

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

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

Wissenschaftliche Texte lesen

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

Wissenschaftlichen Vortrag halten

- 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

Screenshot Web Page

Termine und Themen

Montags um 9:15 Uhr in A.1-13 (office Naumann)

Termin	Thema	Vortrag
22.10.2007	Einführung	Felix Naumann
t.b.d.	Wissenschaftliche Texte Lesen	Felix Naumann
12.11.2007	Literaturkritik / Diskussion	Alle
19.11.2007	Vortragstechniken	Felix Naumann
3.12.2007	Vortrag 1: Mariposa (pdf)	
	Vortrag 2: Trio (pdf)	
10.12.2007	Einführung in LaTeX	Felix Naumann
17.12.2007	Vortrag 3: Fagins Algorithmus (pdf)	
	Vortrag 4: Enough Already in SQL (pdf)	
7.1.2008	Vortrag 5: Sorted Neighborhood (pdf)	
	Vortrag 6: Data Mining (pdf)	
14.1.2008	Vorstellung der Gliederungen & Tipps zur Ausarbeitung	Alle
29.2.2008	Abgabe der Ausarbeitungen	Alle

Überblick

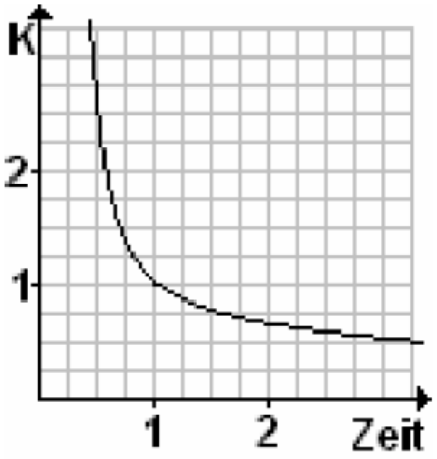
- Vorstellung der Arbeitsgruppe
- Organisatorisches
- Ziele des Seminars
- Zeitlicher Ablauf
- Kurzvorstellung der Themen

Allgemeines

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

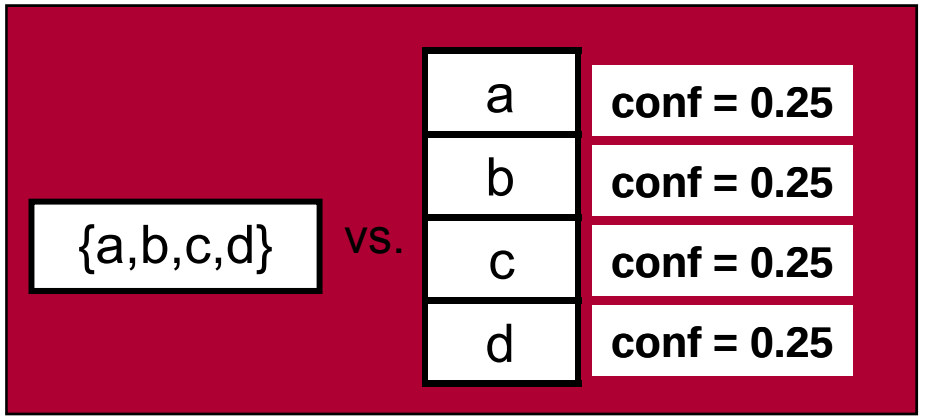
Mariposa

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

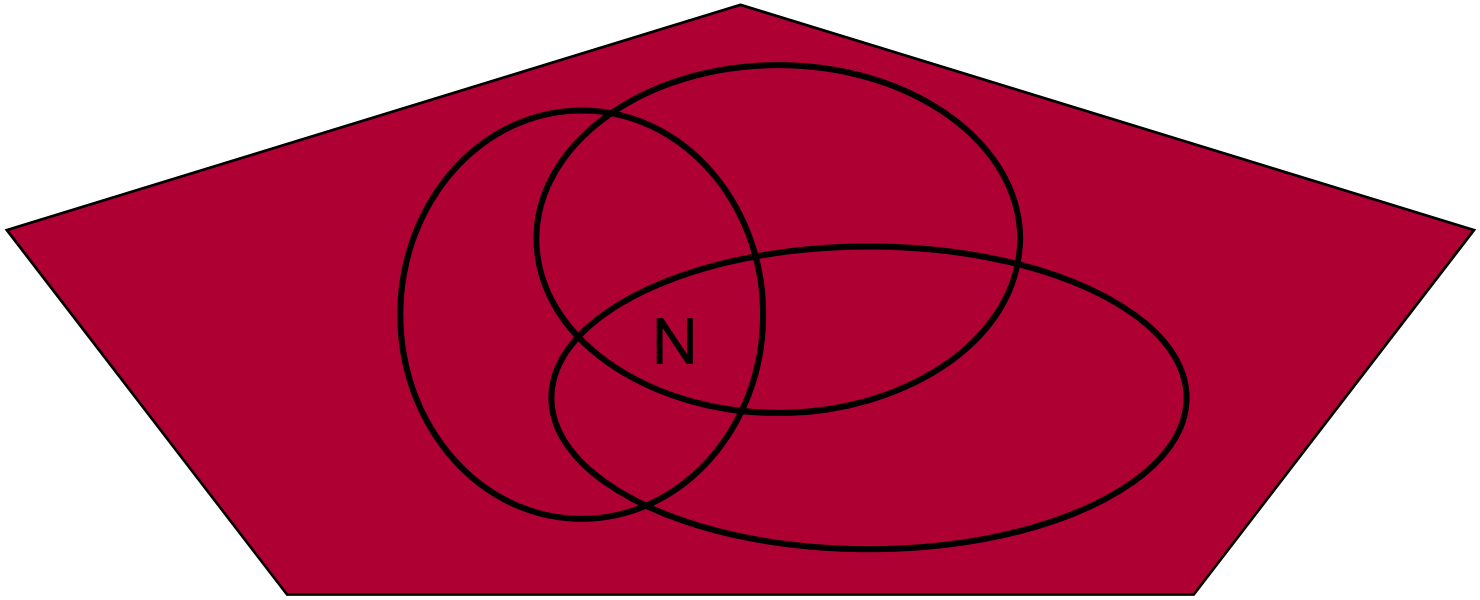


Trio

- DBMS Grundannahmen werden aufgehoben:
  - Jeder Datenwert ist entweder in der DB gespeichert oder nicht.
  - Der Datenwert ist absolut.
  - Seine Herkunft ist irrelevant
- Beispiel: Christmas Bird Count
- Trio relaxiert diese Annahmen:
  - Daten
  - Genauigkeit auf Attributebene, Tupelebene und Relationenebene
  - Herkunft
- Anfragen alle drei Aspekte
  - Schwierige Semantik

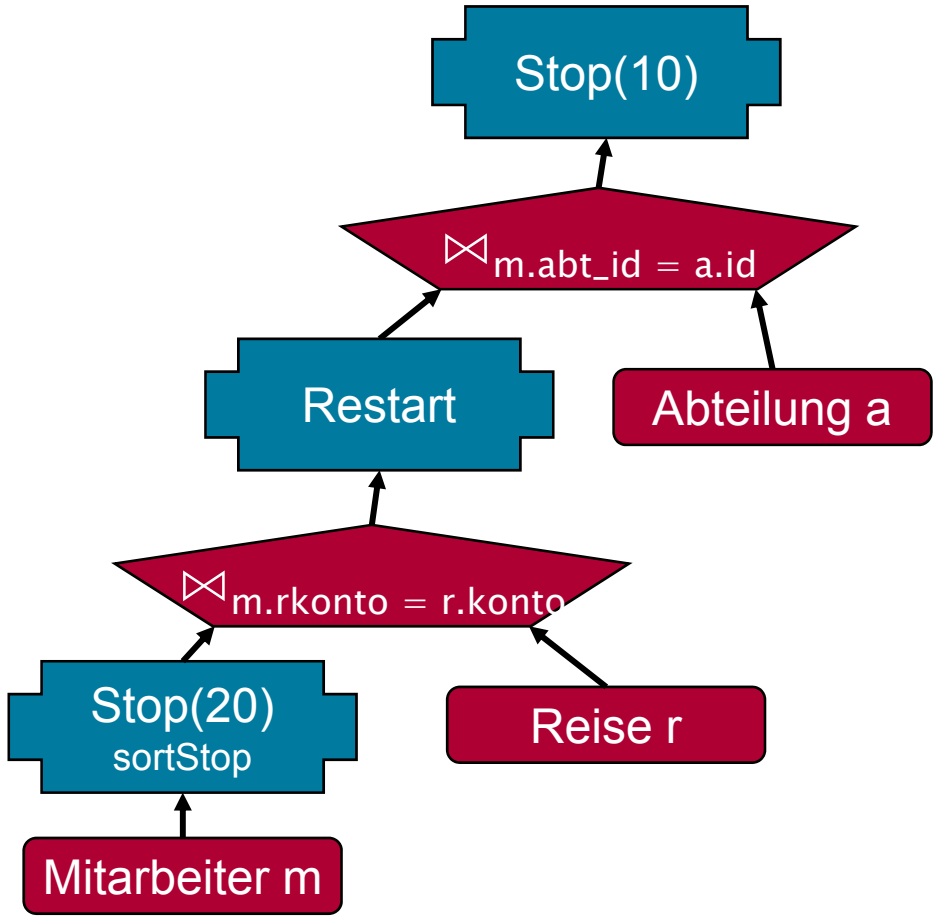


Fagins Algorithmus



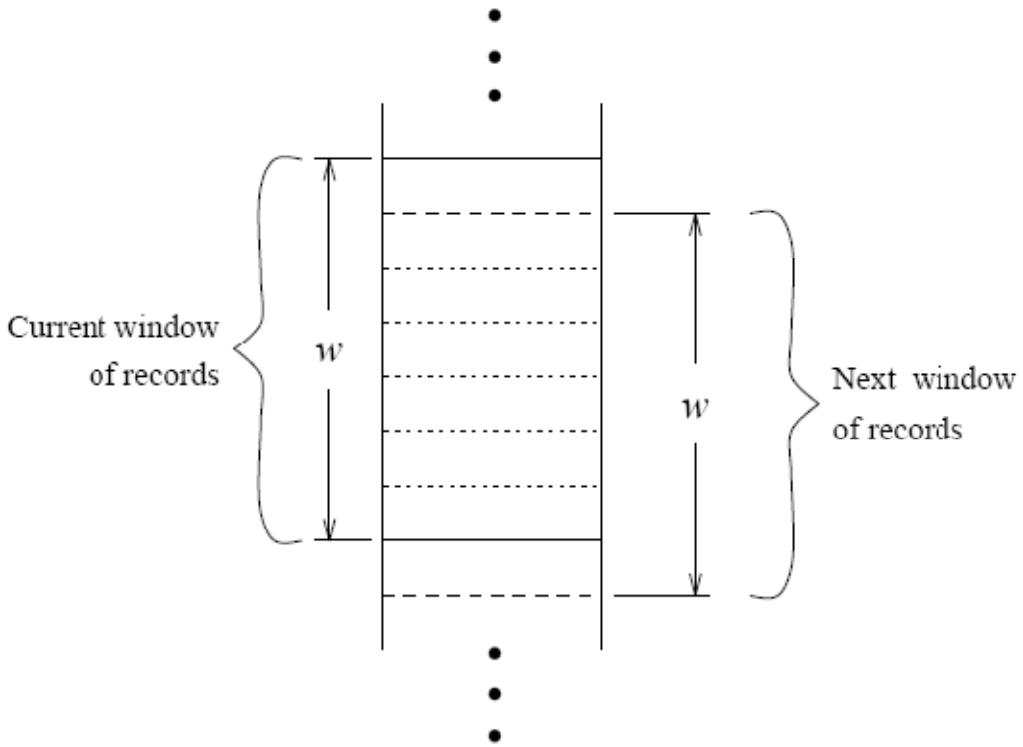
„Enough Already“ in SQL

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



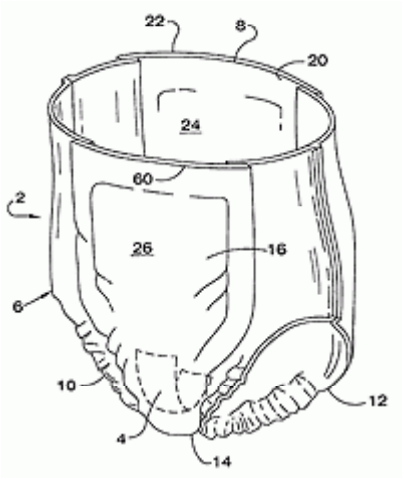
Sorted Neighborhood

- Datenreinigung und Duplikaterkennung
  - Effizienz
  - Effektivität



Data Mining

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



Nächste Schritte

- Sie: E-Mail an mich
  - Mit Wunschliste (Platz 1 – 3)
  - Bis Mitternacht
- Ich: Themenvergabe morgen
  - E-Mail Benachrichtigung
- Ich: Vortrag zum „Lesen wissenschaftlicher Texte“
  - Sondertermin