Information Quality in Integrated Information Systems

Institute for Infocomm Research, Singapore

15.3.2005

Felix Naumann

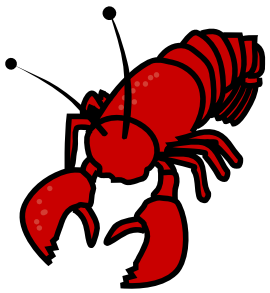
Humboldt-Universität zu Berlin

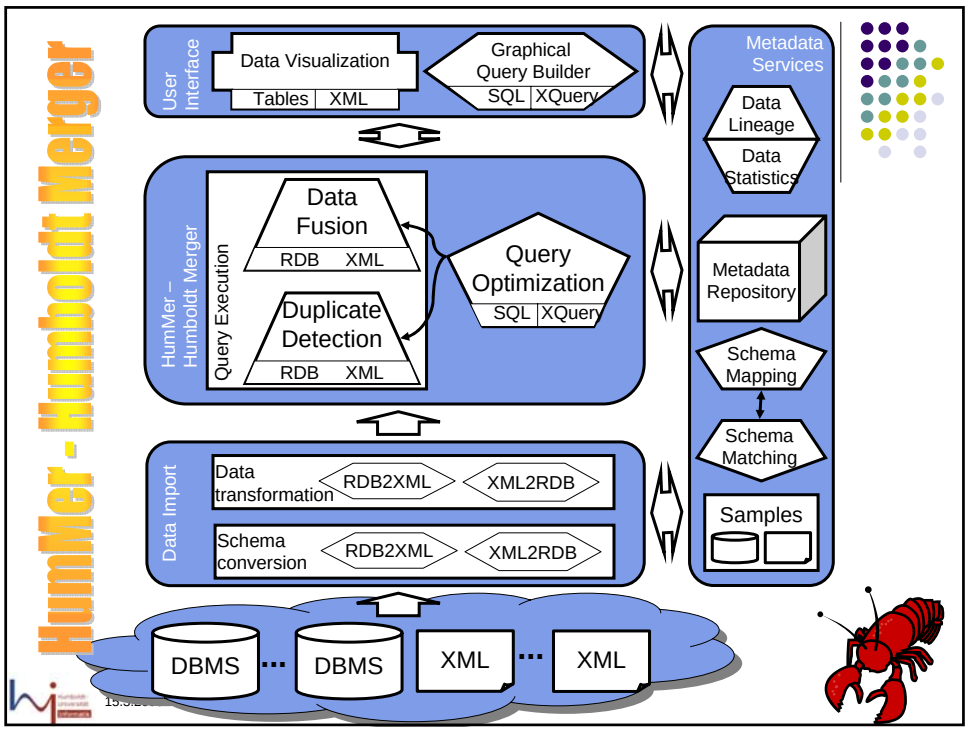
Humboldt-Universität zu Berlin

- Wilhelm and Alexander von Humboldt
- Unity of teaching and research
- Freedom and independence of sciences
- 29 Nobel prize winners
  - Mommsen, Hertz, Koch, Hahn, Planck, Einstein,...
- 38,000 students, (1200 computer sciences)
- 560 professors (12 + 10 computer sciences)

Research Group „Information Integration“

- Head: Felix Naumann (naumann@informatik.hu-berlin.de)
- PhD students
  - Jens Bleiholder (bleiho@informatik.hu-berlin.de)
    - Information fusion in relational data
  - Melanie Weis (mweis@informatik.hu-berlin.de)
    - Object identification in XML Data
- Affiliated PhD students
  - Armin Roth (aroth@informatik.hu-berlin.de)
    - Data quality in Peer-Data-Management-Systems
  - Alexander Bilke (bilke@cs.tu-berlin.de)
    - Schema Matching
- General Research Topics
  - Object identification
  - Information fusion
  - Optimization
  - Visualization
  - And IQ

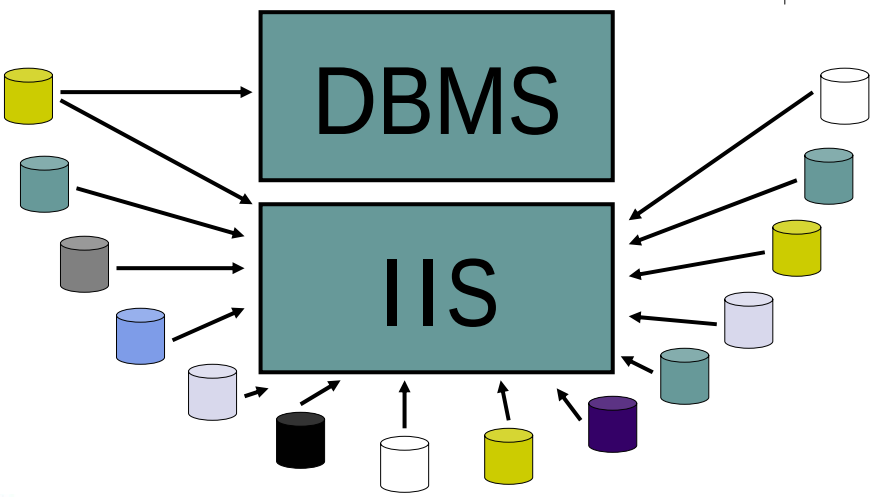




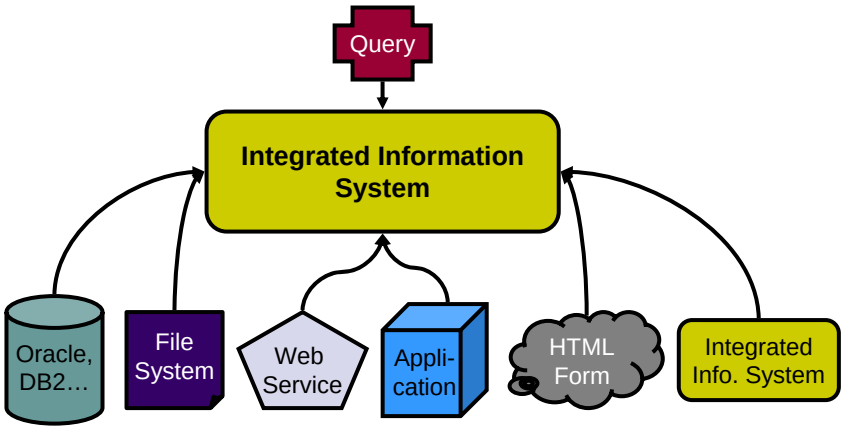
Outline

- Databases vs. Integrated Information systems
- Motivating Examples
- Components for IQ-driven query processing
  - IQ measure + IQ model
  - IQ-driven optimization – A new paradigm
  - IQ-driven integration
- Further topics for discussion

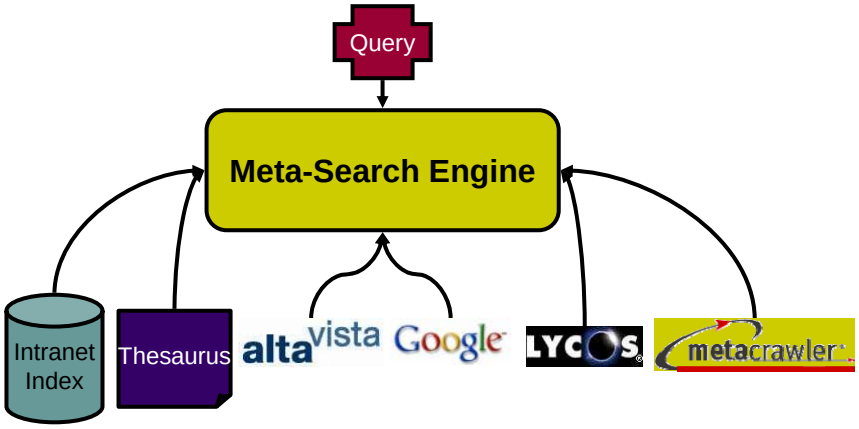
Database Management Systems vs. Integrated Information Systems



Integrated Information Systems

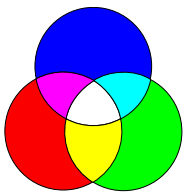


Integrated Search Engines



Redundancy

- Is good for you
  - More information
  - More detailed information
  - Verifiable information
  - Thus: INTEGRATE!



- But yields problems
  - More alternatives mean more complexity.
  - Technical and structural (and semantic) heterogeneity
  - Redundancy is only conceptual:
    - Object identity
    - Data conflicts
  - Mixed and low quality
  - New types of errors (duplicates)

DBMS vs. IIS

Single source	⟺	Multiple sources
Local	⟺	Distributed
Controlled	⟺	Autonomous
Structured	⟺	Semi-structured
Data	⟺	Information
short		text, images, tables
Professional user	⟺	Professional and casual user

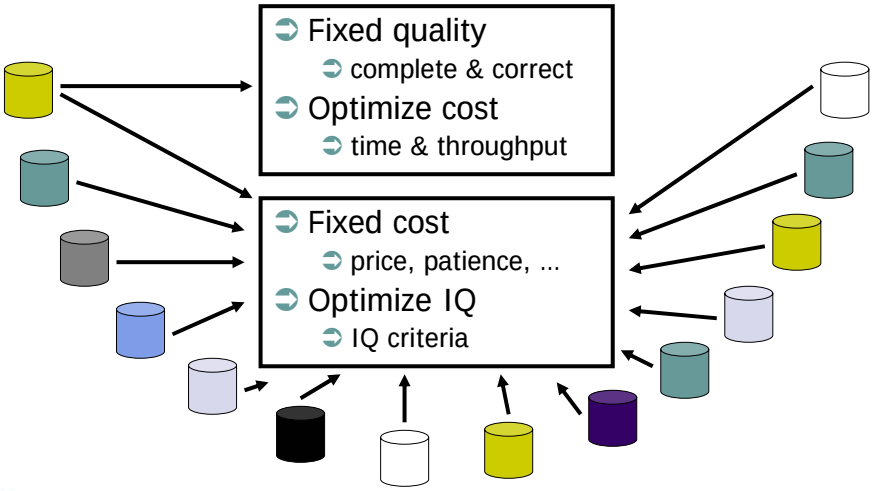
DBMS Quality vs. IIS Quality

Complete (assumed)	⟺	Incomplete
Accurate	⟺	Inaccurate
Trusted	⟺	Untrusted
Fast	⟺	Slow
Free	⟺	Possible cost

High expectations (guarantees) High quality

Low expectations (no guarantees) Low quality

Optimize IQ!



Outline

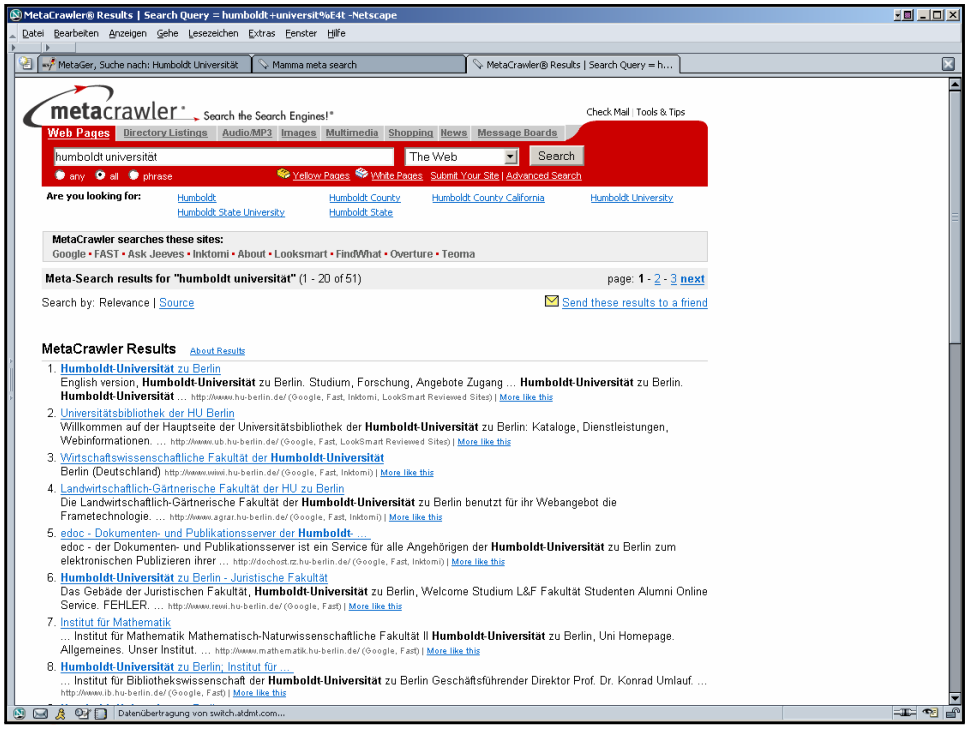
- Databases vs. Integrated Information systems
- Motivating Examples
- Components for IQ-driven query processing
  - IQ measure + IQ model
  - IQ-driven optimization – A new paradigm
  - IQ-driven integration
- Further topics for discussion

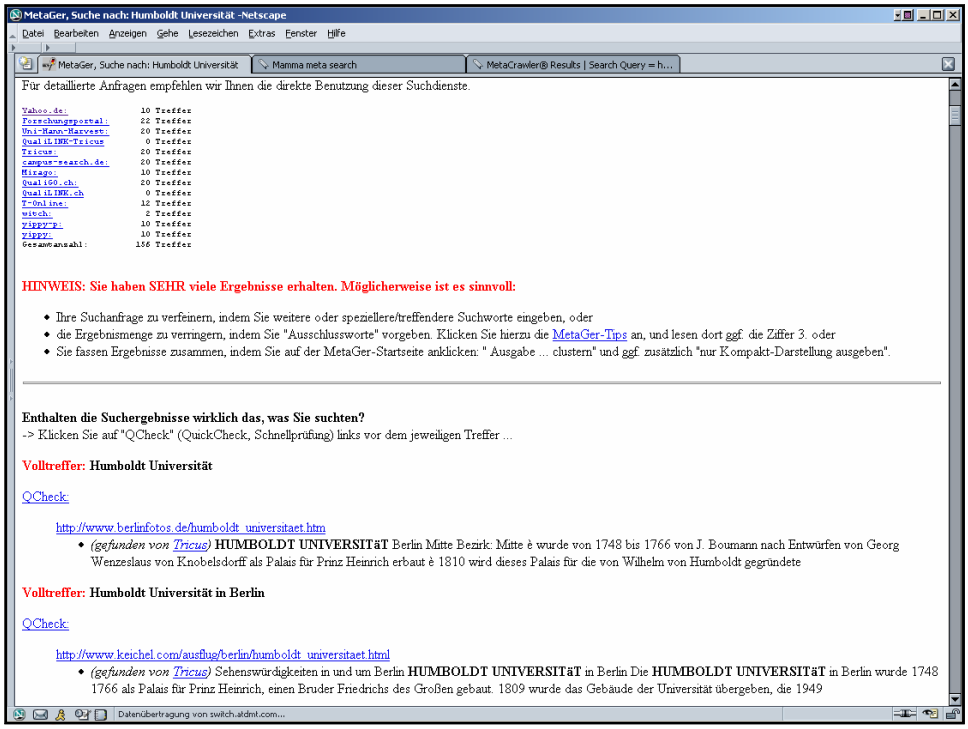
Motivating Examples

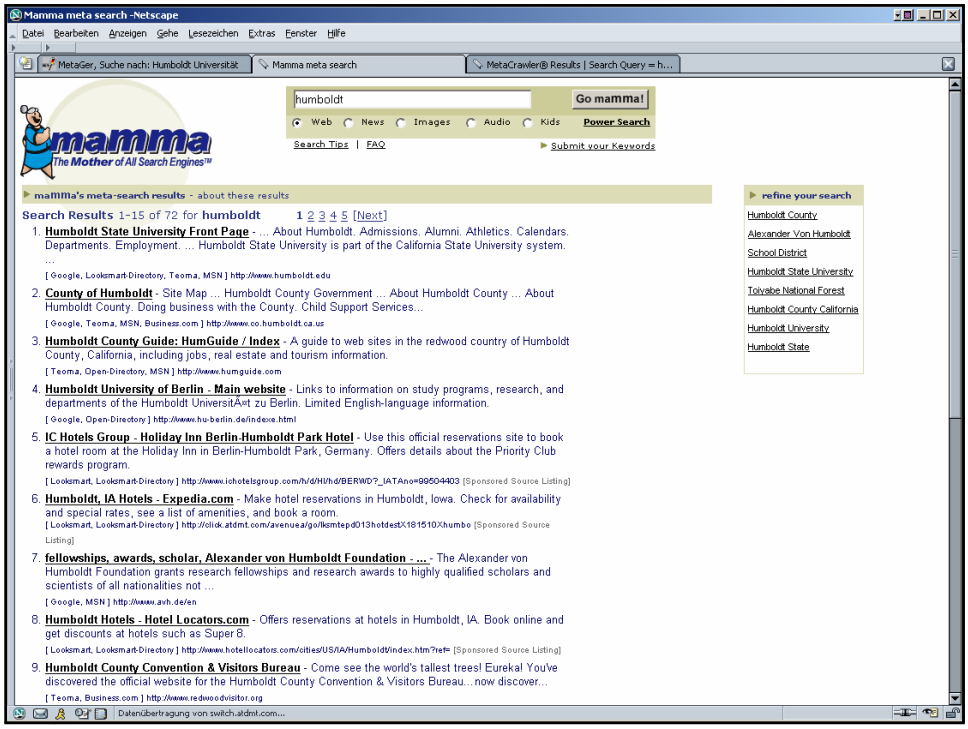
- Search
  - Meta Search Engines
- Data Fusion (Integration)
  - Data conflicts
  - Genome Databases
- Peer Data Management

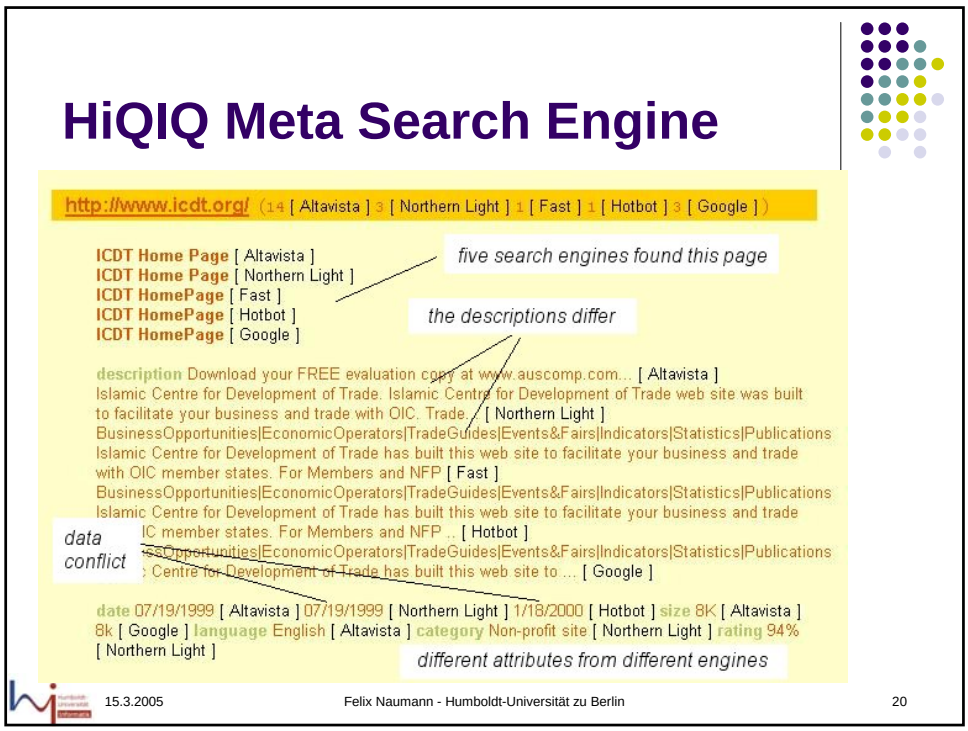
Example: Meta Search Engines

- MetaGer
- Mamma
- MetaCrawler
  - Google: Title, Summary, Descr.,Category, URL, Size
  - Fast: Title, Summary, Descr., URL, Size
  - Inktomi: Title, Summary, URL


MetaCrawler® Results | Search Query = humboldt+universit%E4t -Netscape
Datei Bearbeiten Anzeigen Gehe Lesezeichen Extras Fenster Hilfe
MetaGer, Suche nach: Humboldt Universität
Mamma meta search
MetaCrawler® Results | Search Query = h...
metacrawler
Search the Search Engines!
Check Mail | Tools & Tips
Web Pages
Directory Listings
Audio/MP3
Images
Multimedia
Shopping
News
Message Boards
humboldt universität
The Web
Search
any
all
phrase
Yellow Pages
White Pages
Submit Your Site
Advanced Search
Are you looking for:
Humboldt
Humboldt County
Humboldt County California
Humboldt University
Humboldt State University
Humboldt State
MetaCrawler searches these sites:
Google • FAST • Ask Jeeves • Inktomi • About • Looksmart • FindWhat • Overture • Teoma
Meta-Search results for "humboldt universität" (1 - 20 of 51)
page: 1 - 2 - 3 next
Search by: Relevance | Source
Send these results to a friend
MetaCrawler Results
About Results
1. Humboldt-Universität zu Berlin
English version, Humboldt-Universität zu Berlin. Studium, Forschung, Angebote Zugang ... Humboldt-Universität zu Berlin. Humboldt-Universität ... http://www.hu-berlin.de/ (Google, Fast, Inktomi, LookSmart Reviewed Sites) | More like this
2. Universitätsbibliothek der HU Berlin
Willkommen auf der Hauptseite der Universitätsbibliothek der Humboldt-Universität zu Berlin: Kataloge, Dienstleistungen, Webinformationen. ... http://www.ub.hu-berlin.de/ (Google, Fast, LookSmart Reviewed Sites) | More like this
3. Wirtschaftswissenschaftliche Fakultät der Humboldt-Universität
Berlin (Deutschland) http://www.wiwi.hu-berlin.de/ (Google, Fast, Inktomi) | More like this
4. Landwirtschaftlich-Gärtnerische Fakultät der HU zu Berlin
Die Landwirtschaftlich-Gärtnerische Fakultät der Humboldt-Universität zu Berlin benutzt für ihr Webangebot die Frametechnologie. ... http://www.agrar.hu-berlin.de/ (Google, Fast, Inktomi) | More like this
5. edoc - Dokumenten- und Publikationsserver der Humboldt- ...
edoc - der Dokumenten- und Publikationsserver ist ein Service für alle Angehörigen der Humboldt-Universität zu Berlin zum elektronischen Publizieren ihrer ... http://dochost.rz.hu-berlin.de/ (Google, Fast, Inktomi) | More like this
6. Humboldt-Universität zu Berlin - Juristische Fakultät
Das Gebäde der Juristischen Fakultät, Humboldt-Universität zu Berlin, Welcome Studium L&F Fakultät Studenten Alumni Online Service. FEHLER. ... http://www.rewi.hu-berlin.de/ (Google, Fast) | More like this
7. Institut für Mathematik
... Institut für Mathematik Mathematisch-Naturwissenschaftliche Fakultät II Humboldt-Universität zu Berlin, Uni Homepage. Allgemeines. Unser Institut. ... http://www.mathematik.hu-berlin.de/ (Google, Fast) | More like this
8. Humboldt-Universität zu Berlin, Institut für ...
... Institut für Bibliothekswissenschaft der Humboldt-Universität zu Berlin Geschäftsführender Direktor Prof. Dr. Konrad Umlauf. ... http://www.ib.hu-berlin.de/ (Google, Fast) | More like this
Datenübertragung von switch.atdmt.com...


MetaGer, Suche nach: Humboldt Universität -Netscape
Datei Bearbeiten Anzeigen Gehe Lesezeichen Extras Fenster Hilfe
MetaGer, Suche nach: Humboldt Universität
Mamma meta search
MetaCrawler® Results | Search Query = h...
Für detaillierte Anfragen empfehlen wir Ihnen die direkte Benutzung dieser Suchdienste.
Yahoo.de: 10 Treffer
Forschungsportal: 22 Treffer
Uni*Mann-Harvest: 20 Treffer
Qualil DMK-Tricus 0 Treffer
Tricus: 20 Treffer
campus-search.de: 20 Treffer
Mirago: 10 Treffer
Quali60.ch: 20 Treffer
Qualil DMK.ch 0 Treffer
T-Online: 12 Treffer
witch: 2 Treffer
Zippy-p: 10 Treffer
Zippy: 10 Treffer
Gesamtanzahl: 156 Treffer
HINWEIS: Sie haben SEHR viele Ergebnisse erhalten. Möglicherweise ist es sinnvoll:
Ihre Suchanfrage zu verfeinern, indem Sie weitere oder speziellere/treffendere Suchworte eingeben, oder
die Ergebnismenge zu verringern, indem Sie "Ausschlussworte" vorgeben. Klicken Sie hierzu die MetaGer-Tips an, und lesen dort ggf. die Ziffer 3. oder
Sie fassen Ergebnisse zusammen, indem Sie auf der MetaGer-Startseite anklicken: " Ausgabe ... clustern" und ggf. zusätzlich "nur Kompakt-Darstellung ausgeben".
Enthalten die Suchergebnisse wirklich das, was Sie suchten?
-> Klicken Sie auf "QCheck" (QuickCheck, Schnellprüfung) links vor dem jeweiligen Treffer ...
Volltreffer: Humboldt Universität
QCheck:
http://www.berlinfotos.de/humboldt_universitaet.htm
(gefunden von Tricus) HUMBOLDT UNIVERSITäT Berlin Mitte Bezirk: Mitte è wurde von 1748 bis 1766 von J. Boumann nach Entwürfen von Georg Wenzeslaus von Knobelsdorff als Palais für Prinz Heinrich erbaut è 1810 wird dieses Palais für die von Wilhelm von Humboldt gegründete
Volltreffer: Humboldt Universität in Berlin
QCheck:
http://www.keichel.com/ausflug/berlin/humboldt_universitaet.html
(gefunden von Tricus) Sehenswürdigkeiten in und um Berlin HUMBOLDT UNIVERSITäT in Berlin Die HUMBOLDT UNIVERSITäT in Berlin wurde 1748 1766 als Palais für Prinz Heinrich, einen Bruder Friedrichs des Großen gebaut. 1809 wurde das Gebäude der Universität übergeben, die 1949
Datenübertragung von switch.atdmt.com...

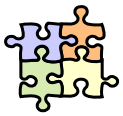

Mamma meta search -Netscape
Datei Bearbeiten Anzeigen Gehe Lesezeichen Extras Fenster Hilfe
MetaGer, Suche nach: Humboldt Universität
Mamma meta search
MetaCrawler® Results | Search Query = h...
mamma
The Mother of All Search Engines™
humboldt
Go mamma!
Web News Images Audio Kids
Power Search
Search Tips | FAQ
Submit your Keywords
mamma's meta-search results - about these results
Search Results 1-15 of 72 for humboldt
1 2 3 4 5 [Next]
1. Humboldt State University Front Page - ... About Humboldt. Admissions. Alumni. Athletics. Calendars. Departments. Employment. ... Humboldt State University is part of the California State University system. ...
[Google, Looksmart-Directory, Teoma, MSN] http://www.humboldt.edu
2. County of Humboldt - Site Map ... Humboldt County Government ... About Humboldt County ... About Humboldt County. Doing business with the County. Child Support Services...
[Google, Teoma, MSN, Business.com] http://www.co.humboldt.ca.us
3. Humboldt County Guide: HumGuide / Index - A guide to web sites in the redwood country of Humboldt County, California, including jobs, real estate and tourism information.
[Teoma, Open-Directory, MSN] http://www.humguide.com
4. Humboldt University of Berlin - Main website - Links to information on study programs, research, and departments of the Humboldt UniversitÄ¤t zu Berlin. Limited English-language information.
[Google, Open-Directory] http://www.hu-berlin.de/indexe.html
5. IC Hotels Group - Holiday Inn Berlin-Humboldt Park Hotel - Use this official reservations site to book a hotel room at the Holiday Inn in Berlin-Humboldt Park, Germany. Offers details about the Priority Club rewards program.
[Looksmart, Looksmart-Directory] http://www.ichotelsgroup.com/h/d/HI/hd/BERWD?_IATAno=99504403 [Sponsored Source Listing]
6. Humboldt, IA Hotels - Expedia.com - Make hotel reservations in Humboldt, Iowa. Check for availability and special rates, see a list of amenities, and book a room.
[Looksmart, Looksmart-Directory] http://click.atdmt.com/avenuea/go/lksmtepd013hotdestX181510Xhumbo [Sponsored Source Listing]
7. fellowships, awards, scholar, Alexander von Humboldt Foundation - ... - The Alexander von Humboldt Foundation grants research fellowships and research awards to highly qualified scholars and scientists of all nationalities not ...
[Google, MSN] http://www.avh.de/en
8. Humboldt Hotels - Hotel Locators.com - Offers reservations at hotels in Humboldt, IA. Book online and get discounts at hotels such as Super 8.
[Looksmart, Looksmart-Directory] http://www.hotellocators.com/cities/US/IA/Humboldt/index.htm?ref= [Sponsored Source Listing]
9. Humboldt County Convention & Visitors Bureau - Come see the world's tallest trees! Eureka! You've discovered the official website for the Humboldt County Convention & Visitors Bureau... now discover...
[Teoma, Business.com] http://www.redwoodvisitor.org
refine your search
Humboldt County
Alexander Von Humboldt
School District
Humboldt State University
Toiyabe National Forest
Humboldt County California
Humboldt University
Humboldt State
Datenübertragung von switch.atdmt.com...


HiQIQ Meta Search Engine
http://www.icdt.org/ (14 [Altavista] 3 [Northern Light] 1 [Fast] 1 [Hotbot] 3 [Google])
ICDT Home Page [Altavista]
ICDT Home Page [Northern Light]
ICDT HomePage [Fast]
ICDT HomePage [Hotbot]
ICDT HomePage [Google]
five search engines found this page
the descriptions differ
description Download your FREE evaluation copy at www.auscomp.com... [Altavista]
Islamic Centre for Development of Trade. Islamic Centre for Development of Trade web site was built to facilitate your business and trade with OIC. Trade... [Northern Light]
BusinessOpportunities|EconomicOperators|TradeGuides|Events&Fairs|Indicators|Statistics|Publications
Islamic Centre for Development of Trade has built this web site to facilitate your business and trade with OIC member states. For Members and NFP [Fast]
BusinessOpportunities|EconomicOperators|TradeGuides|Events&Fairs|Indicators|Statistics|Publications
Islamic Centre for Development of Trade has built this web site to facilitate your business and trade
data conflict
IC member states. For Members and NFP .. [Hotbot]
ssOpportunities|EconomicOperators|TradeGuides|Events&Fairs|Indicators|Statistics|Publications
Centre for Development of Trade has built this web site to ... [Google]
date 07/19/1999 [Altavista] 07/19/1999 [Northern Light] 1/18/2000 [Hotbot] size 8K [Altavista]
8k [Google] language English [Altavista] category Non-profit site [Northern Light] rating 94% [Northern Light]
different attributes from different engines
15.3.2005
Felix Naumann - Humboldt-Universität zu Berlin
20

Motivating Examples

- Search
  - Meta Search Engines
- Data Fusion (Integration)
  - Data conflicts
  - Genome Databases
- Peer Data Management

Example: Data Fusion



Source 1 → a, b
Source 2 → a, b
=> a, b

a, b
a, -
=> a, b

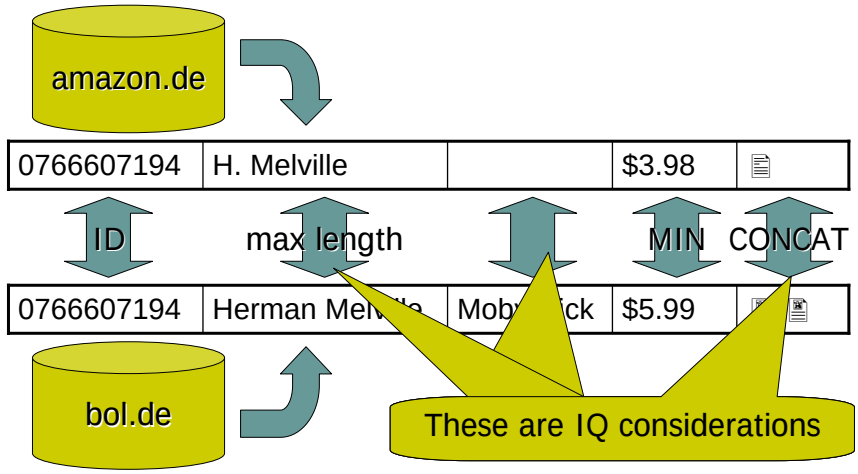
Meaning of null?

a, b
a, c
=> a, r(b,c)

Conflict-Resolution

a, -, d
a, c, -
=> a, c, d

Example: Data Fusion



Example: Data Fusion

- Genome Databases
  - Research labs produce many conceptually redundant results.
  - Quality of experimental results vary (poor IQ is intrinsic).
  - Quality of human (and machine) annotations vary.
  - Biologists are very opinionated about IQ of information sources.
  - ⇒ Perfect application for IQ-driven information integration

Example: Data Fusion

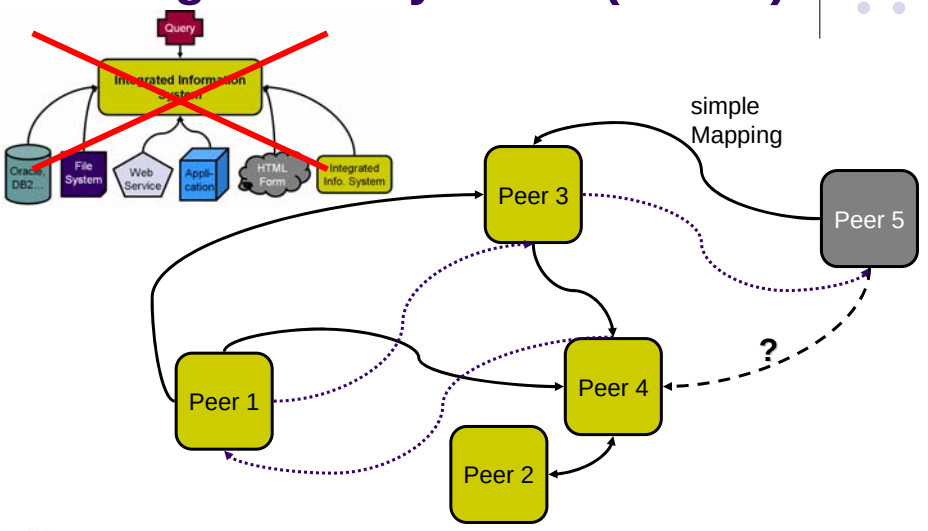
- Genome databases
  - Identification of redundancy very difficult (no standard representation)
  - No IQ metadata,
  - only human-readable annotation
  - (Some annotation on “methods used”)

RNA sequence entry from EMBL

Motivating Examples

- Search
  - Meta Search Engines
  - Information Retrieval
- Data Fusion (Integration)
  - Data conflicts
  - Genome Databases
- Peer Data Management

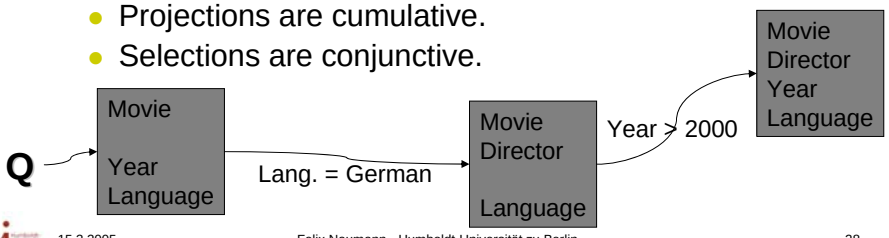
Example: Peer Data Management Systems (PDMS)



Work with Armin Roth, DaimlerChrysler

Example: Peer Data Management Systems (PDMS)

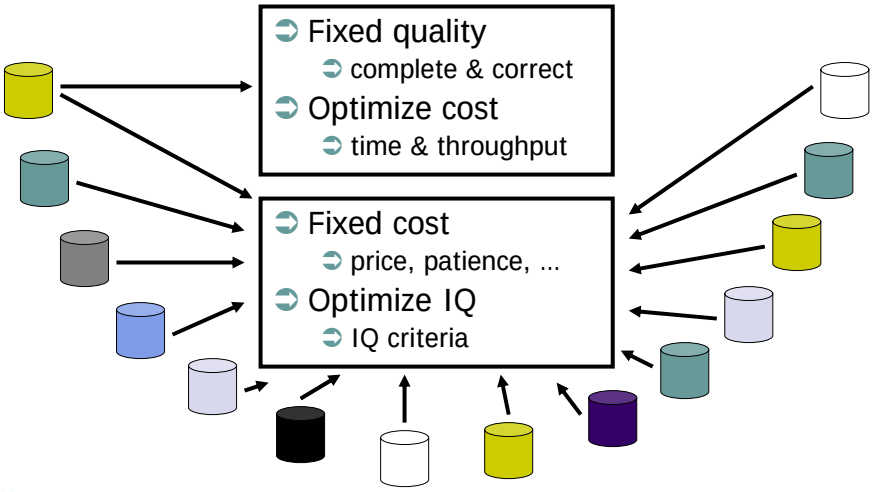
- IQ in PDMS
  - You don‘t know who is answering your query.
  - Mappings further degrade IQ:
    - Projections diminish density
    - Selections diminish completeness
  - Problem compounded along mapping paths:
    - Projections are cumulative.
    - Selections are conjunctive.



Outline

- Databases vs. Integrated Information systems
- Motivating Examples
- Components for IQ-driven query processing
  - IQ measure + IQ model
  - IQ-driven optimization – A new paradigm
  - IQ-driven integration
- Further topics for discussion

Optimize IQ!

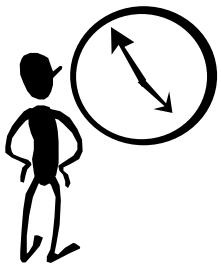


A New Optimization Paradigm – Changes

- Cost criteria ⟺ Quality criteria
- Cost model ⟺ Quality model
- Optimization algorithm ⟺ Optimization algorithm
- + Information integration

DB Cost Criteria

- Response time
- Execution time
- Latency
- Throughput
- Cardinality
- …



Assessed through system parameters and statistics.

IIS Quality Criteria

IQ :=

> ***"Even though quality cannot be defined, you know what it is."***
>
> **Robert Pirsig**

IIS Quality Criteria

IQ := {Understandability, Reputation, Reliability, Timeliness, Availability, Price, Conciseness Consistency, Coverage, Response time, Density, Completeness, Amount, Accuracy, Relevancy, ... }

IQ Criteria

- Accuracy
  - Definition:
    - Usually: Percentage of incorrect tuples
    - For integration: Percentage of incorrect data values
  - Assessment:
    - Domain and Constraint Testing
    - Lookup tables
    - Scientific measurements
    - Data-input experience
  - Improvement:
    - Often: Deletion
    - Better: “Data Scrubbing”
- Response Time
  - Definition:
    - Usually: Time until complete query result is received
    - For integration: Latency
  - Assessment:
    - “Cost Calibration”
    - Continuous assessment
  - Improvement:
    - Source selection
    - Classical optimization
    - Federated Optimization

IQ Criteria

- Completeness
  - Definition:
    - Coverage (extension): Number of real world objects represented
    - Density (intension): Number of attributes covered
      - NULL-values
  - Assessment:
    - Sampling
    - Existing Metadata
  - Improvement:
    - Source selection
    - “Best k” vs. “k best”
- Conciseness
  - Definition:
    - Extension: Number of real-world objects not represented doubly
    - Intension: Number of attributes not represented doubly
  - Assessment:
    - Duplicate detection
    - Schema Matching
  - Improvement:
    - Data Fusion

IIS Quality Criteria

IQ := {Understandability, Reputation, Reliability, Timeliness, Availability, Price, Conciseness Consistency, Coverage, Response time, Density, Completeness, Amount, Accuracy, Relevancy, ... **}**

Assessed in 3 classes...

IQ Assessment

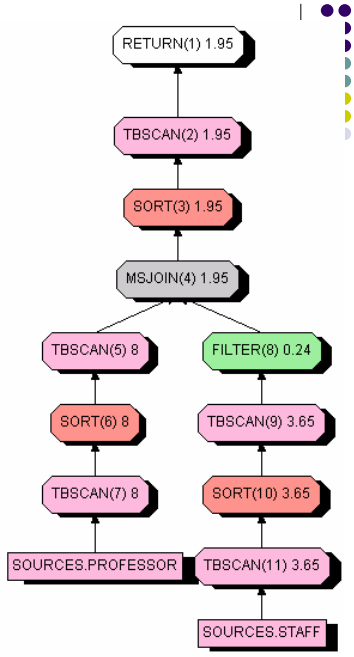
Information source:	S_1	S_2	S_3
Understandability	5	7	7
Reputation	5	5	7
Reliability	2	6	4
Age	30	30	2
Availability	99	99	60
Completeness	1	1	0.5
Response Time	0.2	0.2	0.2
Accuracy	99.9	99.9	99.8
Relevancy	60	80	90

- User
  - Questionnaires
- Information Source
  - Parsing
  - Metadata
- Query Process
  - Time Measurement

DB Cost Models

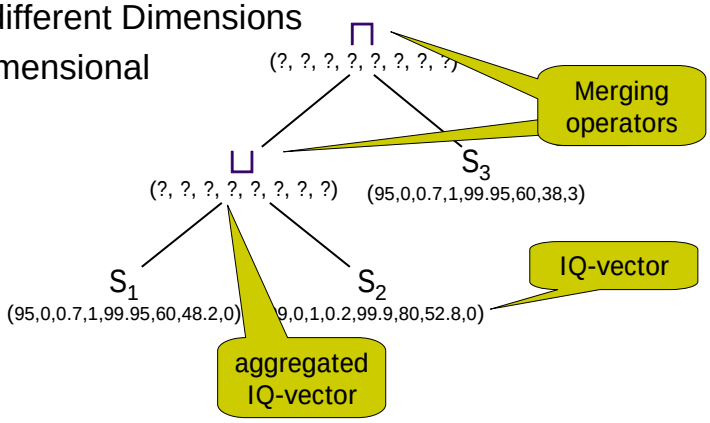
- Operators
  - + (add)
  - max
  - X (multiply)

Easy ;-)



IIS Quality Model

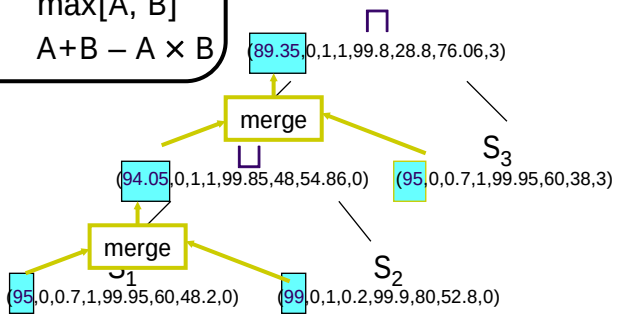
- 2 Problems
  - Many different Dimensions
  - Multidimensional



Merge IQ in many Dimensions

IQ Merge Functions

Availability:	A × B
Price:	A + B
Response Time:	max[A, B]
Coverage:	A+B – A × B



Multidimensional IQ

(89.35,0,1,1,99.8,28.8,76.06,3) > (82.35,0,2,1.5,95,32,71.77,2) ?

- IQ-criteria have
  - Different units
  - Different ranges
  - Different importance
- So...
  - convert
  - scale
  - weight

MADM methods:
SAW, TOPSIS, ELECTRE, AHP, DEA

Outline

- Databases vs. Integrated Information systems
- Motivating Examples
- Components for IQ-driven query processing
  - IQ measure + IQ model
  - IQ-driven optimization – A new paradigm
  - IQ-driven integration
- Further topics for discussion

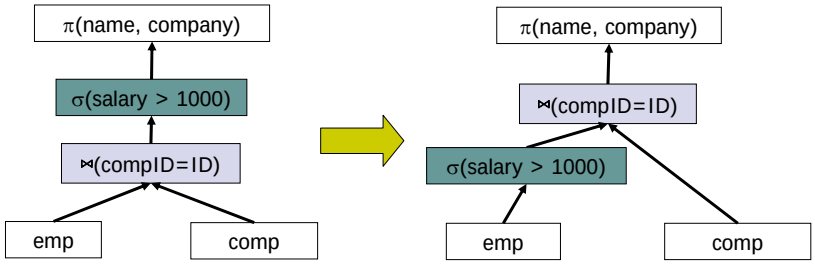
DB-type Optimization

- Goal
  - Minimize response time (single user)
  - Maximize throughput (multi-user)
- Restrictions
  - Complete
  - Correct (not just accurate: filter conditions…)

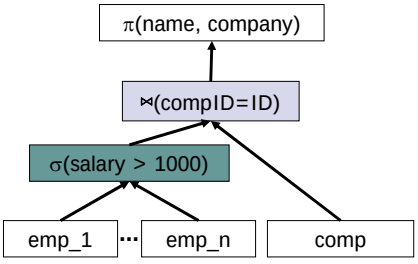
Find best plan!

DB-type Optimization

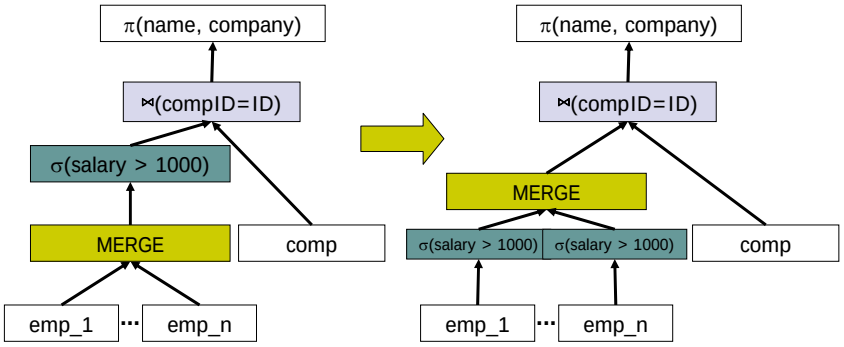
```
SELECT name, company FROM emp, comp
WHERE emp.compID = comp.ID AND emp.salary > 1000
```



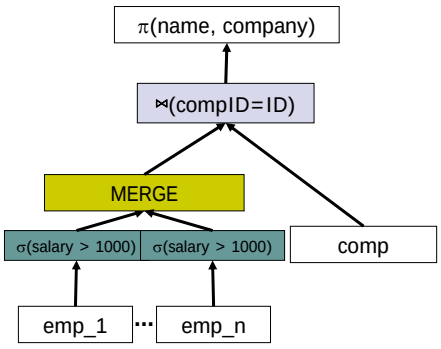
DB-type Optimization



Integration-type Optimization



Integration-type Optimization



- Change is efficient
- But:
  - Result can be incomplete.
  - Depending on “Merge function”
- Preferences? Quality!

IIS-type Optimization

- Goal
  - Maximize information quality
  - Maximize completeness
- Restrictions
  - Price
  - Bandwidth
  - Time (user patience)

Find *K* best sources – Find best *K* sources

IIS-type Optimization

- *K* best sources
  - (or *K* best conjunctive plans)
  - Simple IQ model, but
  - Sources may not complement each other
    - at tuple level (replication)
    - at attribute level
- Best *K* sources
  - (or best disjunctive normal form plan)
  - Finds optimal query result
  - Uses IQ merging

Outline

- Databases vs. Integrated Information systems
- Motivating Examples
- Components for IQ-driven query processing
  - IQ measure + IQ model
  - IQ-driven optimization – A new paradigm
  - IQ-driven integration
- Further topics for discussion

Data Integration in DBMS

- UNION, UNION, UNION
- OUTER UNION,
- GROUP BY

- Cannot compute similarity.
  - For object identification
- Cannot deal with inconsistencies.
  - For conflict resolution

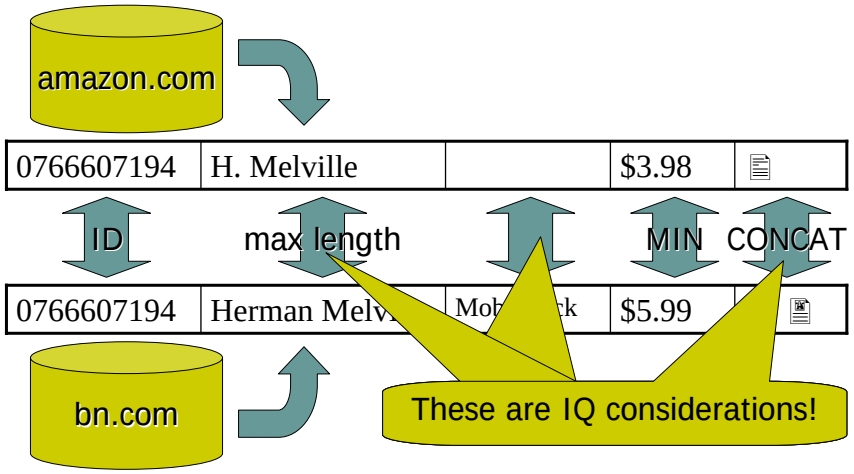
Information Integration in IS

- WHAT?
  - Problem: Identify different representations of same real-world object.
  - Solution: Object identification techniques
- HOW?
  - Problem: Merge conflicting data about same real-world object.
  - Solution: Conflict resolution techniques

Object Identification

- Also
  - Duplicate Detection
  - Record Linkage
  - Data Reconciliation/Consolidation
- Domain-specific
  - Address data
  - Microbiological data
  - …
- IQ weighting can help

Conflict Resolution



Conflict Resolution

- Numerical:
  SUM, AVG, MAX, MIN, …
- Non-numerical:
  MAXLENGTH, CONCAT, AnnCONCAT,…
- Special:
  RANDOM, COUNT, CHOOSE, **FAVOR**, **MaxIQ**,…
- Domain-specific
  …

Human specification of IQ

Automated specification of IQ

Outline

- Databases vs. Integrated Information systems
- Motivating Examples
- Components for IQ-driven query processing
  - IQ measure + IQ model
  - IQ-driven optimization – A new paradigm
  - IQ-driven integration
- Further topics for discussion

Further topics for discussion

- One slide each on ongoing work:
  - IQ in Off-the-Shelf-DBMS
  - Detecting regular errors in DBMS using Data Mining
  - Duplicate detection in XML
  - Duplicate detection without a schema
  - Data Fusion – Completeness and Conciseness
  - 10th Int. Conference on Information Quality

IQ in Off-the-Shelf-DBMS

1. Most quality-critical data is stored in DBMS.
   - Data Warehouses
   - Web back-ends
2. DBMS are carefully designed to preserve high data quality.
   - Decades of development
   - ACID
3. Sometimes DBMS even improve data quality.
   - Aggregation
   - Integration

Work with Mary Roth, IBM

Detecting regular errors in DBMS using Data Mining

r_1

ID	ORG.	SEX	COLOR	SIZE
1	Frog	F	Green	20,4
2	Toad	F	Green	25,1
3	Newt	M	Grey	30,4
4	Frog	F	Blue	19,6
5	Toad	M	Blue	20,8
6	Newt	M	Grey	30,1
7	Newt	F	Grey	28,9
8	Frog	M	Green	20,3

r_2

ID	ORG.	SEX	COLOR	SIZE
1	Frog	W	Olive	20
2	Frog	W	Green	25,1
3	Newt	M	Grey-Spotted	30,4
4	Frog	W	Blue	19,6
5	Frog	M	Blue	20,8
6	Newt	M	Grey-Spotted	30,1
7	Newt	W	Grey&Yellow	28,9
8	Frog	M	Olive	20

r_m

			C[CORGANISM]			C[SEX]			C[COLOR]			C[SIZE]
ID	R_1.ORG	R_2.ORG	C_1	R_1.SEX	R_2.SEX	C_2	R_1.COLOR	R_2.COLOR	C_3	R_1.SIZE	R_2.SIZE	C_4
1	Frog	Frog	0	F	W	1	Green	Olive	1	20,4	20	1
2	Toad	Frog	1	F	W	1	Green	Green	0	25,1	25,1	0
3	Newt	Newt	0	M	M	0	Grey	Grey-Spotted	1	30,4	30,4	0
4	Frog	Frog	0	F	W	1	Blue	Blue	0	19,6	19,6	0
5	Toad	Frog	1	M	M	0	Blue	Blue	0	20,8	20,8	0
6	Newt	Newt	0	M	M	0	Grey	Grey-Spotted	1	30,1	30,1	0
7	Newt	Newt	0	F	W	1	Grey	Grey&Yellow	1	28,9	28,9	0
8	Frog	Frog	0	M	M	0	Green					

Work with Heiko Müller, Humboldt

Duplicate detection in XML

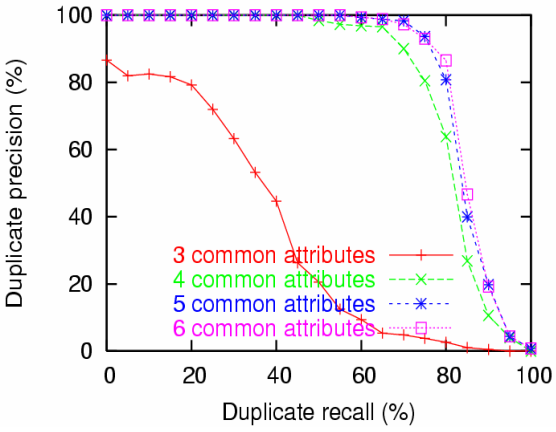
```
- <author>
    <name>Bernd Amann</name>
  - <publication>
      <title>XML Repository and Active Views Demonstration.</title>
      <year>1999</year>
    </publication>
  - <publication>
      <title>Active Views for Electronic Commerce.</title>
      <year>1999</year>
    </publication>
  </author>
- <author>
    <name>Sophie Cluet</name>
  - <publication>
      <title>XML Repository and Active Views Demonstration.</title>
      <year>1999</year>
    </publication>
  - <publication>
      <title>Active Views for Electronic Commerce.</title>
      <year>1999</year>
    </publication>
  - <publication>
      <title>Views in a Large Scale XML Repository.</title>
      <year>2001</year>
    </publication>
  </author>
```

- Compare <author>
  - With Subelements (<publication>)?
  - How deep?
- Compare <publication>
  - With siblings (<year>)?
  - Schema, or Data?
- In short: What is a duplicate?

Work with Melanie Weis, Humboldt

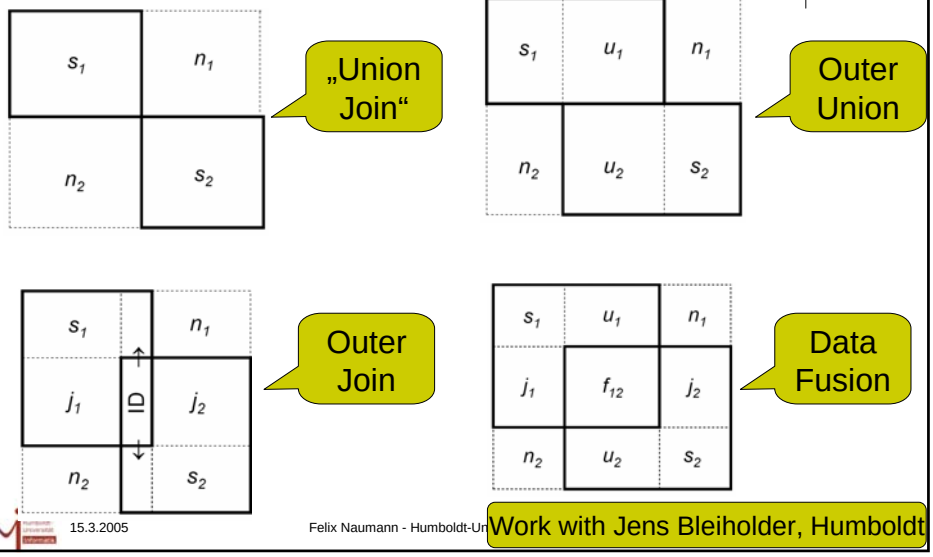
Duplicate detection without a schema

- Opaque attribute names
- Only some intensional overlap
- Similarity measure based on tokenization of tuple and TFIDF
- Whirl algorithm to efficiently find top duplicates



Work with Alexander Bilke, University of Technology Berlin

Data Fusion – Completeness and Conciseness



Work with Jens Bleiholder, Humboldt

CfP: 10th ICIQ

- International Conference on Information Quality
- Nov. 4 to 6 @ MIT in Cambridge
- Paper deadline: June 30 !
- Suggested topics:
  - IQ Concepts, Tools, Metrics, Measures, Models, and
  - Methodologies
  - IQ Policies and Standards
  - IQ Assessment
  - IQ Practices: Case Studies and Experience Reports
  - IQ Product Experience Reports
  - Cost/Benefit Analysis of IQ and IQ Improvement
  - Information Product Implementation, Delivery, and
  - Management
  - IQ in Databases, the Web, and e-Business
  - Data Warehouses and Data Mining
  - Corporate Household Data
  - IQ in Scientific Data Management
  - Data Cleansing and Reconciliation
  - IQ Education and Curriculum Development
  - Trust, Knowledge, and Society in the IQ Context
- Academic, research-in-progress, or practice-oriented paper.

Questions

- Databases vs. Integrated Information systems
- Motivating Examples
- Components for IQ-driven query processing
  - IQ measure + IQ model
  - IQ-driven optimization – A new paradigm
  - IQ-driven integration

www.hiqiq.de
www.informatik.hu-berlin.de/mac/
naumann@informatik.hu-berlin.de