Information Quality: Fundamentals, Techniques, and Use

Felix Naumann
Humboldt-Universität zu Berlin

Kai-Uwe Sattler
TU Ilmenau

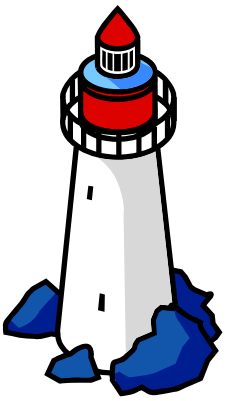
EDBT Tutorial, Munich, March 28 2006

Our Personal Motivation

- Now: Motivation
  - IQ is big business
  - IQ is (also) a database topic
- This tutorial: The past
  - Where we are now
- The future: Open Problems
  - Much to do

1.5 hours ⇒ no details

Tutorial Overview



- Motivation
- Defining IQ
  - IQ Dimensions
  - IQ Models
- IQ Assessment
  - Assessment techniques
  - IQ aggregation and ranking
- IQ Improvement
  - Profiling & Data Scrubbing
  - Outlier Detection
  - Duplicate Detection
- Wrapup

Information Quality: Fundamentals, Techniques, and Use

Part 1: Motivation

Felix Naumann
Humboldt-Universität zu Berlin

Kai-Uwe Sattler
TU Ilmenau

EDBT Tutorial, Munich, March 28 2006

Anecdotal Evidence

- Incorrect prices in inventory retail databases [English 1999]
  - Costs for consumers 2.5 billion $
  - 80% of barcode-scan-errors to the disadvantage of consumer
- IRS 1992: almost 100,000 tax refunds not deliverable [English 1999]
- 50% to 80% of computerized criminal records in the U.S. were found to be inaccurate, incomplete, or ambiguous. [Strong et al. 1997a]
- US-Postal Service: of 100,000 mass-mailings up to 7,000 undeliverable due to incorrect addresses [Pierce 2004]

Goodyear reveals $100 million error 10/23/07 USA

Goodyear said late Wednesday that it will restate earnings for the past five years, decreasing income by as much as $100 million because an accounting system caused billing errors. The tiremaker is delaying the release of its third-quarter earnings, expected this morning, until mid-November. Shares closed up 2 cents to $6.83 before the announcement; in after-hours trading, shares plummeted 27%, or $1.83, to $5.

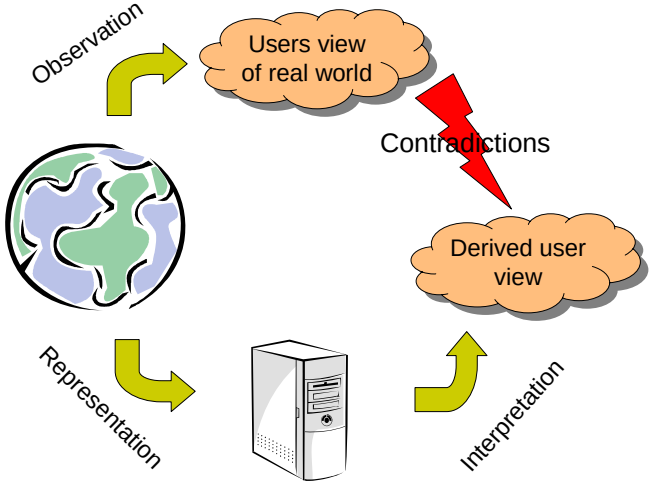
Anecdotal Evidence

- In 2006 the Fortune 1000 companies will spend more money on IQ problems than for ERP, CRM, and BI together. [Gartner]
- More than 35% of all IT projects fail due to poor IQ. Poor IQ causes annual expenses of 2-4 billion $ in US. [Meta Group]
- IQ is one of the most important success factors in DWH and CRM projects. [PriceWaterhouseCoopers]
- Data collection in the wake of 2004 tsunami
  - Fatalities and injuries
  - Housing damages
  - Property damages
- http://www.informationquality.org/publiclyexposediqproblems.cfm

Examples

- Data warehousing (DWH)
  - Incorrect price of item in inventory table
    - 10,000 $ ≠ 10.000 $
  - 2000 orders for this item
    - Revenue 20,000,000 $ or 20,000 $?
  - Decision: Increase marketing?
- Customer relationship management (CRM)
  - Revenue with **Bayerische Motoren Werke AG** 1,000,000 €
  - Revenue with **BMW** 60,000 €
  - Revenue with **BaMoWe** 15,000,000 €
  - Question: Is BMW a preferred customer?
- Further examples: Healthcare, Disaster data management, etc
- Materialized and virtual integration

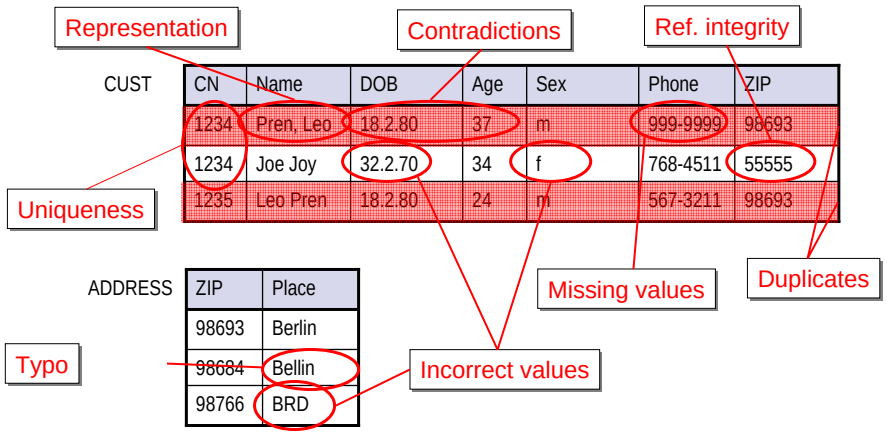
Causes of Poor Information Quality



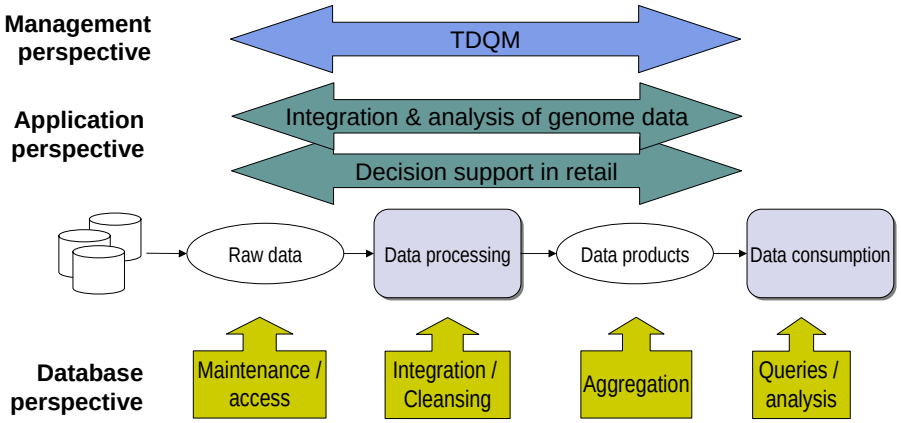
Causes of Poor Information Quality

- Data production
  - Data collection with human input (typos etc.)
  - Systematic problems with data collection (Incorrect codes, etc.)
  - Different sources with different representations of same real world object
- Storage
  - Different formats
  - Insufficient formats
- Usage
  - Insufficient analysis and processing capabilities
  - Change of IQ requirements
  - Security and access problems

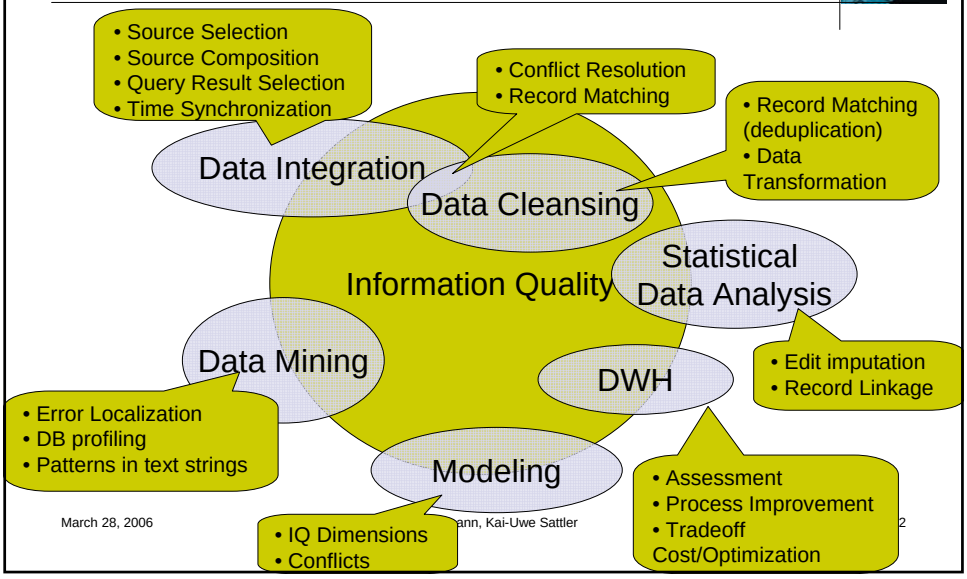
Information Quality Problems



Perspectives



Much to be done [Batini et al. 2004]



References

- [Batini et al. 2004]
  C. Batini, T. Catarci, M. Scannapieco: *A Survey of Data Quality Issues in Cooperative Information Systems*, Int. Conference on Conceptual Modeling (ER 2004), Shanghai, China, 2004.
- [English 1999]
  L. English: *Improving Data Warehouse and Business Information Quality*, Wiley, 1999.
- [Pierce 2004]
  E. Pierce: *Assessing Data Quality with Control Matrices*, Communications of the ACM 47(2): 82-86, 2004.
- [Strong et al. 1997a]
  D. Strong, Y. Lee, R. Wang: *Data Quality in Context*, Communications of the ACM 40(5): 103-110, 1997.

Information Quality: Fundamentals, Techniques, and Use

Part 2: Defining IQ

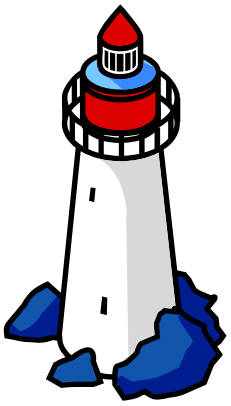
Felix Naumann
Humboldt-Universität zu Berlin

Kai-Uwe Sattler
TU Ilmenau

EDBT Tutorial, Munich, March 28 2006

Overview

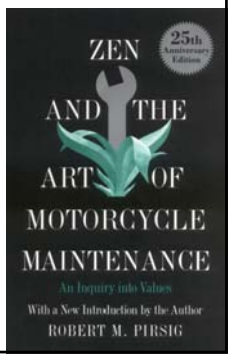
- Motivation
- Defining IQ
  - Dimensions and Classifications
  - Models
- IQ Assessment
- IQ Improvement
- Wrapup



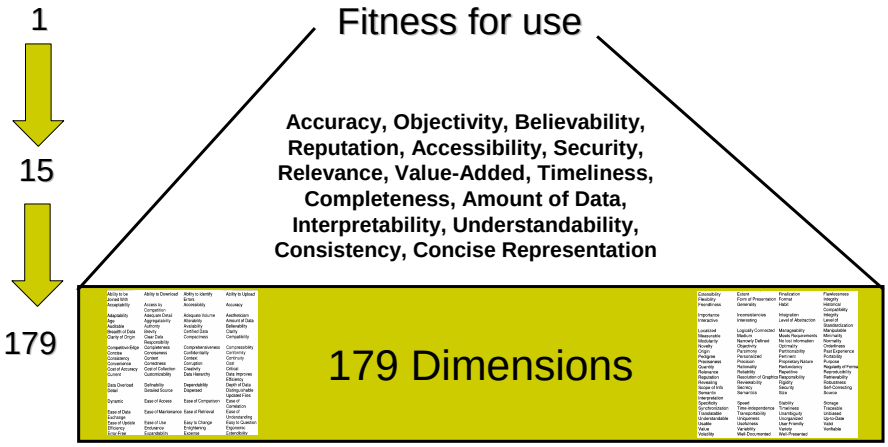
Quality

"Even though quality cannot be defined, you know what it is."

Robert Pirsig



Zooming into IQ



IQ from 10000 feet

- General definitions
  - „excellence / value“
  - „fitness for use“
  - „extent to which a product successfully serves the purpose of consumers“
  - „meeting / exceeding consumer expectations“
  - „inexact science in terms of assessment and benchmarks“
- Observations
  - Information quality is **subjective**
    - Depends on context, consumer, etc.
  - Information quality is **multidimensional**
    - multiple dimensions (criteria, aspects, properties)

IQ under the Microscope

Ability to be Joined With	Ability to Download	Ability to Identify Errors	Ability to Upload
Acceptability	Access by Competition	Accessibility	Accuracy
Adaptability	Adequate Detail	Adequate Volume	Aestheticism
Age	Aggregatability	Alterability	Amount of Data
Auditable	Authority	Availability	Believability
Breadth of Data	Brevity	Certified Data	Clarity
Clarity of Origin	Clear Data Responsibility	Compactness	Compatibility
Competitive Edge	Completeness	Comprehensiveness	Compressibility
Concise	Conciseness	Confidentiality	Conformity
Consistency	Content	Context	Continuity
Convenience	Correctness	Corruption	Cost
Cost of Accuracy	Cost of Collection	Creativity	Critical
Current	Customizability	Data Hierarchy	Data Improves Efficiency
Data Overload	Definability	Dependability	Depth of Data
Detail	Detailed Source	Dispersed	Distinguishable
			Updated Files
Dynamic	Ease of Access	Ease of Comparison	Ease of Correlation
Ease of Data Exchange	Ease of Maintenance	Ease of Retrieval	Ease of Understanding
Ease of Update	Ease of Use	Easy to Change	Easy to Question
Efficiency	Endurance	Enlightening	Ergonomic
Error-Free	Expandability	Expense	Extendibility

[Wang Strong 1996]

IQ under the Microscope

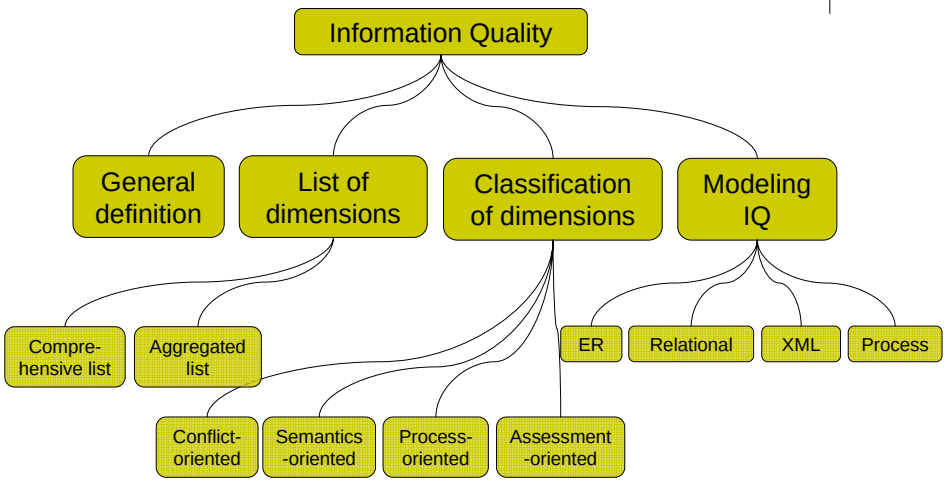
Extensibility	Extent	Finalization	Flawlessness
Flexibility	Form of Presentation	Format	Integrity
Friendliness	Generality	Habit	Historical Compatibility
Importance	Inconsistencies	Integration	Integrity
Interactive	Interesting	Level of Abstraction	Level of Standardization
Localized	Logically Connected	Manageability	Manipulable
Measurable	Medium	Meets Requirements	Minimality
Modularity	Narrowly Defined	No lost information	Normality
Novelty	Objectivity	Optimality	Orderliness
Origin	Parsimony	Partitionability	Past Experience
Pedigree	Personalized	Pertinent	Portability
Preciseness	Precision	Proprietary Nature	Purpose
Quantity	Rationality	Redundancy	Regularity of Forma
Relevance	Reliability	Repetitive	Reproducibility
Reputation	Resolution of Graphics	Responsibility	Retrievability
Revealing	Reviewability	Rigidity	Robustness
Scope of Info	Secrecy	Security	Self-Correcting
Semantic Interpretation	Semantics	Size	Source
Specificity	Speed	Stability	Storage
Synchronization	Time-independence	Timeliness	Traceable
Translatable	Transportability	Unambiguity	Unbiased
Understandable	Uniqueness	Unorganized	Up-to-Date
Usable	Usefulness	User Friendly	Valid
Value	Variability	Variety	Verifiable
Volatility	Well-Documented	Well-Presented	

[Wang Strong 1996]

Finding the right Dimensions

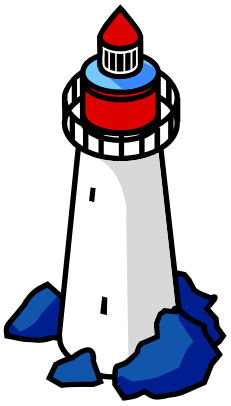
IQ := {Understandability, Reputation, Reliability, Timeliness, Availability, Price, Consistency, Coverage, Response time, Density, Completeness, Amount, Accuracy, Relevancy, ... **}**

Defining IQ



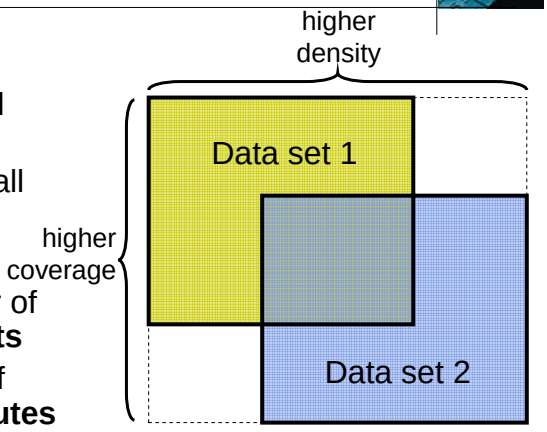
Overview

- Motivation
- Defining IQ
  - Dimensions and Classifications
  - Models
- IQ Assessment
- IQ Improvement
- Wrapup



Selected IQ Dimensions – Completeness

- Completeness
  - Fraction of non-null values
  - Representation of all real-world objects
- Also
  - Coverage: Number of represented **objects**
  - Density: Number of represented **attributes**
  - [Naumann et al. 2004]



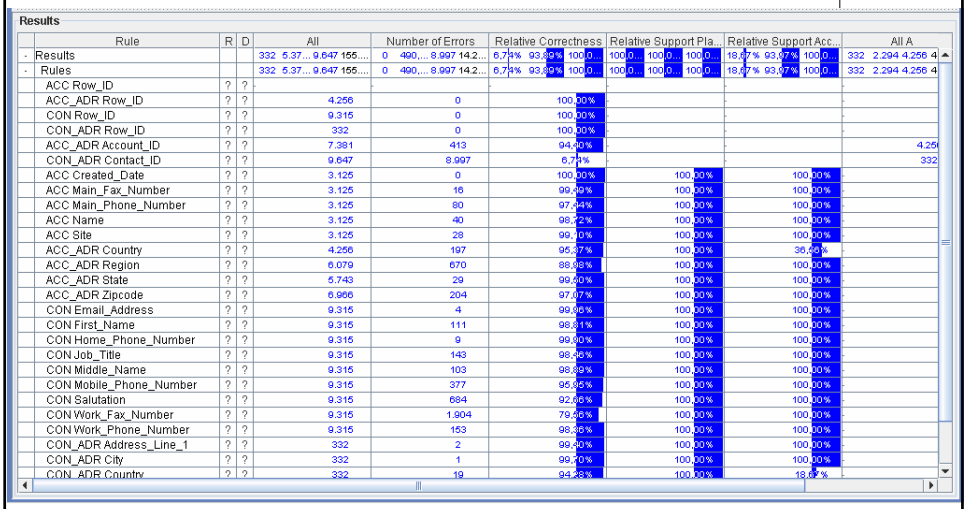
Selected IQ Dimensions – Completeness

- The extent to which data are of sufficient **breadth**, **depth** and **scope** for the task at hand
  - [Wang Strong 1996]
- Coverage denotes the estimated **portion** of the **intended** complete **relation** that is actually present.
  - Trio System [Widom 2005]
- A subset of a database is complete if it includes a representation of **every occurrence** in the real world environment that it models.
  - [Motro 1986]
- Soundness measures the proportion of the stored information that is true, and completeness measures the **proportion of the true information** that is stored.
  - [Motro Rakov 1998]
- Coverage is the **probability** that a source has some answer to a given query.
  - [Florescu et al. 1997]

Selected IQ Dimensions – Accuracy

- The extent to which data are correct, reliable, and certified free of error.
  - [Wang Strong 1996]
- At value level: Fraction of correct values
  - In a tuple
  - In a relation
- At tuple level: Fraction of tuples with only correct values
- Definition of correctness
  - Nearness of a value to the correct value
  - Missing values?
  - Rules, domains, etc.

Example Accuracy Report



Results

Rule	R	D	All	Number of Errors	Relative Correctness	Relative Support Pla...	Relative Support Acc...	All A
Results			332 5.37... 9.647 155....	0 490,... 8.997 14.2...	6,7...% 93,3...% 100,0...	100,0... 100,0... 100,0...	18,67% 93,07% 100,0...	332 2.294 4.256 4
Rules			332 5.37... 9.647 155....	0 490,... 8.997 14.2...	6,7...% 93,3...% 100,0...	100,0... 100,0... 100,0...	18,67% 93,07% 100,0...	332 2.294 4.256 4
ACC Row_ID	?	?						
ACC_ADR Row_ID	?	?	4.256	0	100,00%			
CON Row_ID	?	?	9.315	0	100,00%			
CON_ADR Row_ID	?	?	332	0	100,00%			
ACC_ADR Account_ID	?	?	7.381	413	94,40%			4.25
CON_ADR Contact_ID	?	?	9.647	8.997	6,7...%			332
ACC Created_Date	?	?	3.125	0	100,00%	100,00%	100,00%	
ACC Main_Fax_Number	?	?	3.125	16	99,49%	100,00%	100,00%	
ACC Main_Phone_Number	?	?	3.125	80	97,44%	100,00%	100,00%	
ACC Name	?	?	3.125	40	98,72%	100,00%	100,00%	
ACC Site	?	?	3.125	28	99,10%	100,00%	100,00%	
ACC_ADR Country	?	?	4.256	197	95,37%	100,00%	36,00%	
ACC_ADR Region	?	?	6.079	670	88,98%	100,00%	100,00%	
ACC_ADR State	?	?	5.743	29	99,50%	100,00%	100,00%	
ACC_ADR Zipcode	?	?	6.966	204	97,07%	100,00%	100,00%	
CON Email_Address	?	?	9.315	4	99,96%	100,00%	100,00%	
CON First_Name	?	?	9.315	111	98,81%	100,00%	100,00%	
CON Home_Phone_Number	?	?	9.315	9	99,90%	100,00%	100,00%	
CON Job_Title	?	?	9.315	143	98,46%	100,00%	100,00%	
CON Middle_Name	?	?	9.315	103	98,89%	100,00%	100,00%	
CON Mobile_Phone_Number	?	?	9.315	377	95,95%	100,00%	100,00%	
CON Salutation	?	?	9.315	684	92,66%	100,00%	100,00%	
CON Work_Fax_Number	?	?	9.315	1.904	79,56%	100,00%	100,00%	
CON Work_Phone_Number	?	?	9.315	153	98,36%	100,00%	100,00%	
CON_ADR Address_Line_1	?	?	332	2	99,40%	100,00%	100,00%	
CON_ADR City	?	?	332	1	99,70%	100,00%	100,00%	
CON_ADR Country	?	?	332	19	94,28%	100,00%	18,67%	

More IQ Dimensions

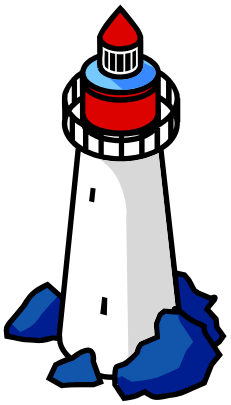
- Content-related dimensions (data-intrinsic)
  - Accuracy, completeness, relevancy, interpretability, value-added
- Technical dimensions (hard- and software)
  - Reliability, response time, latency, QoS, price, security, timeliness
- Intellectual dimensions (subjective)
  - Believability, reputation/trust, objectivity
- Instance-related dimensions (presentation)
  - Amount of data, understandability, concise and consistent representation, verifiability

Many Classifications

- Classification based on conflicts
  - [Rahm Do 2000] and [Redman 1996]
  - Schema vs. data
  - One source vs. multiple sources
- Semantic classifications
  - [Naumann 2002] and [Strong et al. 1997a]
- Process-oriented classification
  - [Liu Chi 2002]
- Classification for IQ assessment
  - [Naumann Rolker 2000]

Overview

- Motivation
- Defining IQ
  - Dimensions and Classifications
  - Models
- IQ Assessment
- IQ Improvement
- Wrapup

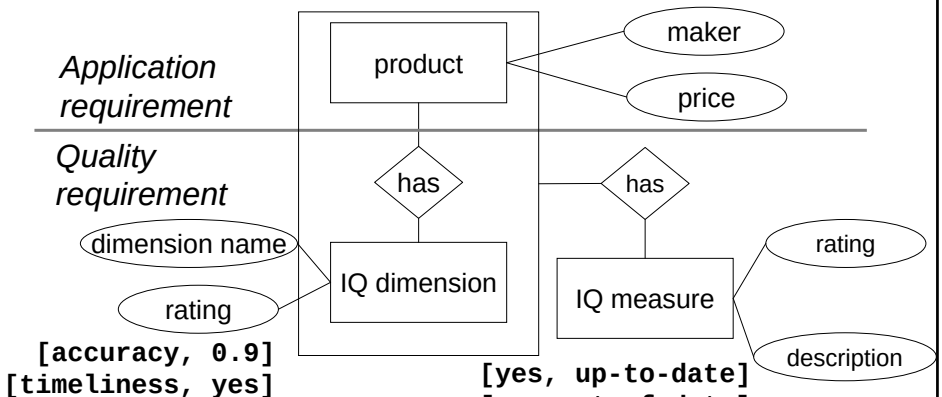


IQ Models

- Data models
  - Common theme: Enrich conventional data model with elements to represent and analyze IQ.
  - Conceptional modeling
    - ER-Extension: Quality ER-Model [Storey Wang 1998]
  - Logical modeling
    - Extension of relational model
      - Polygen [Wang Madnick 1990]
      - Attribute-based model [Wang et al. 1995]
    - Trio DBMS for data, accuracy, and lineage [Widom 2005]
    - Extended XML-Model: D2Q [Scannapieco et al. 2004]
- Process model
  - Model for data production process
    - IP-MAP [Shankaranarayanan et al. 2000, Wang et al. 2003]

IQ Representation in ER Model

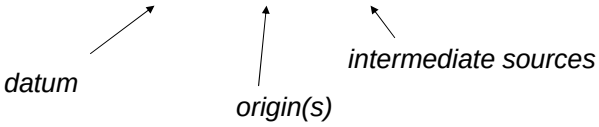
„*Most data models (including the ER model) assume that all data given is correct.*“ [Chen 1993]



[Storey Wang 1998]

Polygen

- Explicit representation of origin (lineage)
- Extension of the relational model
- Attribute value in a Polygen relation is a triplet
  (c<d>, c<o>, c<i>)

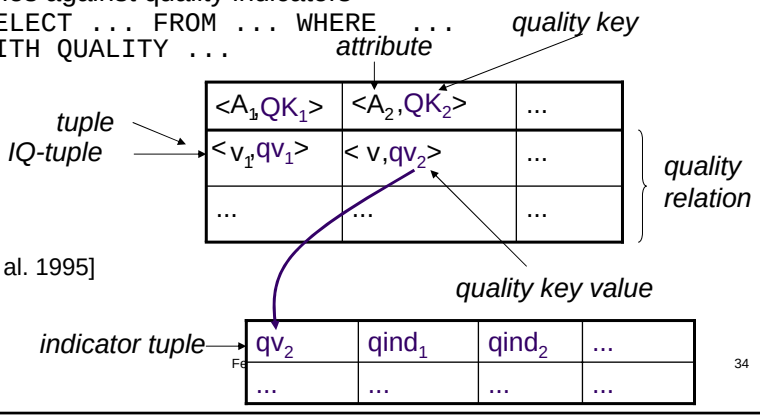


- Extension of relational operators (projection, selection, etc.) to also update intermediate sources

[Wang Madnick 1990]

Attribute-based Model

- Extension of the relational model: „cell tagging“
  - Attributes with different levels of quality indicators
- Extension of the relational algebra
- Queries against quality indicators
  - `SELECT ... FROM ... WHERE ... WITH QUALITY ...`



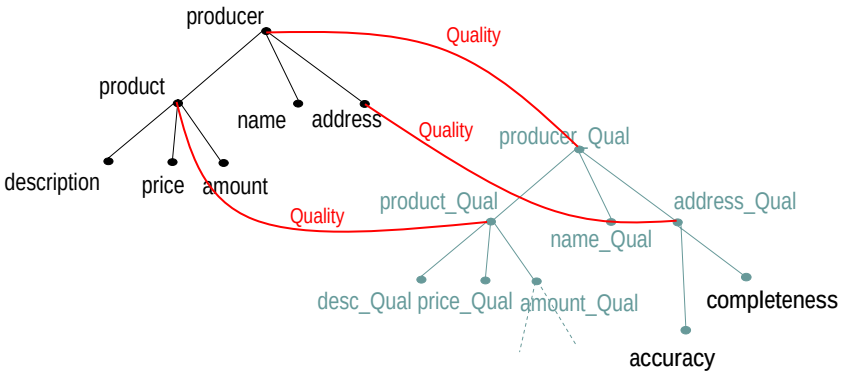
[Wang et al. 1995]

Trio

- A system for integrated management of
  1. Data
  2. Accuracy
     - Attribute-level: approximation
     - Tuple-level (or relation-level): confidence
     - Relation-level: coverage
  3. Lineage
- Data Model (triples)
- Algebra for relational operators
- Query Language: TriQL

[Widom 2005]

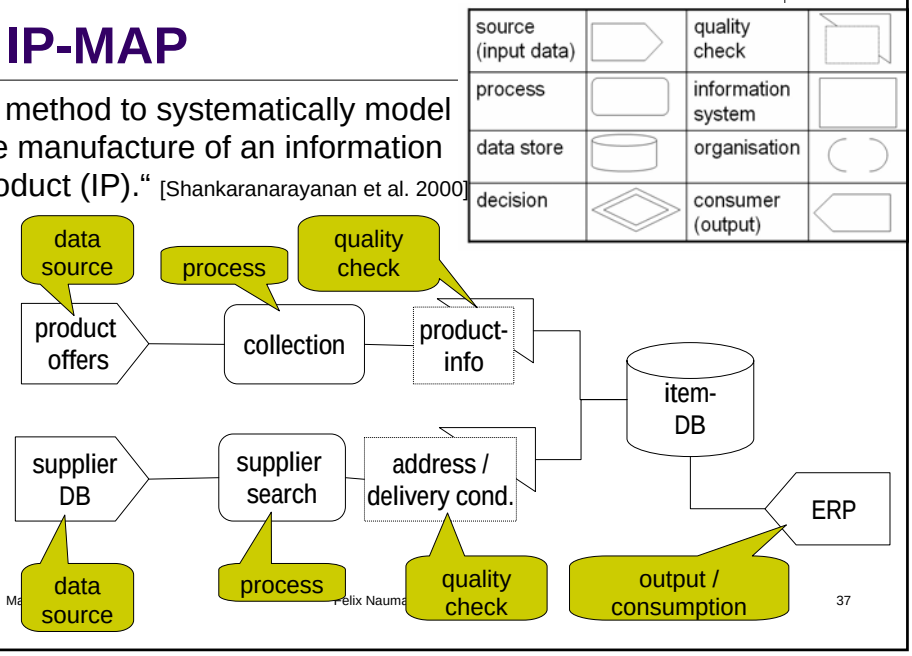
Data and Data Quality (D²Q)



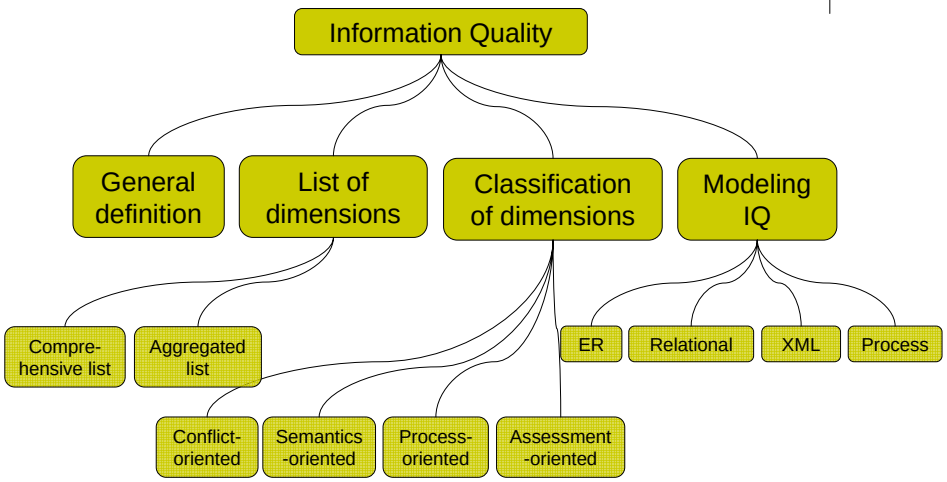
[Scannapieco et al. 2004]

IP-MAP

„A method to systematically model the manufacture of an information product (IP).“ [Shankaranarayanan et al. 2000]



Summary – IQ Dimensions and Models



References

- [Chen 1993]
  Peter Chen: *The Entity-Relationship Approach* in Information Technology in Action: Trends and Perspectives. R.Y. Wang, editor, Prentice Hall 1993
- [Florescu et al. 1997]
  Daniela Florescu, Daphne Koller, Alon Y. Levy: *Using Probabilistic Information in Data Integration*. VLDB 1997: 216-225
- [Liu Chi 2002]
  L. Liu, L. Chi: *Evolutional Data Quality: A Theory-specific View*, Proc. of the Int. Conference on Information Quality (IQ 2002), pages 292-304, 2002.
- [Motro 1986]
  Amihai Motro: *Completeness Information and Its Application to Query Processing*. VLDB 1986: 170-178
- [Motro Rakov 1998]
  A. Motro, I. Rakov: *Estimating the Quality of Databases*, Proc. of the Int. Conference on Flexible Query Answering (FQAS 1998), pages 298-307, 1998.
- [Naumann et al. 2004]
  Felix Naumann, Johann Christoph Freytag, Ulf Leser: *Completeness of integrated information sources*. Inf. Syst. 29(7): 583-615 (2004)
- [Naumann Rolker 2000]
  Felix Naumann, Claudia Rolker: *Assessment Methods for Information Quality Criteria*. IQ 2000: 148-162
- [Naumann 2002]
  Felix Naumann: *Quality-Driven Query Answering for Integrated Information Systems,* Springer 2002
- [Rahm Do 2000]
  Erhard Rahm, Hong Hai Do: *Data Cleaning: Problems and Current Approaches*. IEEE Data Eng. Bull. 23(4): 3-13 (2000)
- [Redman 1996]
  T. Redman: *Data Quality for the Information Age*, Artech House, 1996.

References

- [Storey Wang 1998]
  V. Storey, R. Wang: *An Analysis of Quality Requirements in Database Design*, Proc. of the Int. Conference on Information Quality (IQ 1998), pages 64-87, 1998.
- [Strong et al. 1997a]
  D. Strong, Y. Lee, R. Wang: *Data Quality in Context*, Communications of the ACM 40(5): 103-110, 1997.
- [Scannapieco et al. 2004]
  M. Scannapieco, A. Virgillito, C. Marchetti, M. Mecella, R. Baldoni: *The DaQuinCIS Architecture: A Platform for Exchanging and Improving Data Quality in Cooperative Information Systems*, Information Systems 29(7): 551-582, 2004.
- [Shankaranarayanan et al. 2000]
  G. Shankaranarayanan, R. Wang, M. Ziad: *IP-MAP: Representing the Manufacture of an Information Product*, Proc. of the Int. Conference on Information Quality (IQ 2000), pages 1-16, 2000.
- [Wang et al. 2003]
  R. Wang, T. Allen, W. Harris, S. Madnick: *An Information Product Approach for Total Information Awareness*, Proc. of IEEE Aerospace Conference, 2003.
- [Wang Madnick 1990]
  R. Wang, S. Madnick: *A Polygen Model for Heterogeneous Database Systems: The Source Tagging Perspective*, Proc. of the 16th VLDB Conference, Brisbane, Australia, pages 519-538, 1990.
- [Wang et al. 1995]
  R. Wang, M. Reddy, H. Kon: *Towards Quality Data: An Attribute-based Approach*, Decision Support Systems 13:349-372, 1995.
- [Wang Strong 1996]
  R. Wang and D. Strong: *Beyond Accuracy: What Data Quality Means to Data Consumers*, Journal of Management Information Systems, Vol. 12, No. 4, 1996, 5-34
- [Widom 2005]
  Jennifer Widom: *Trio: A System for Integrated Management of Data, Accuracy, and Lineage*. CIDR 2005: 262-276

Information Quality: Fundamentals, Techniques, and Use

Part 3: Assessment

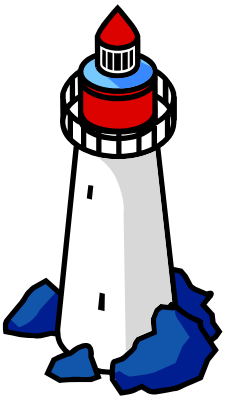
Felix Naumann
Humboldt-Universität zu Berlin

Kai-Uwe Sattler
TU Ilmenau

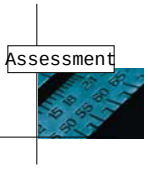
EDBT Tutorial, Munich, March 28 2006

Tutorial Overview

- Motivation
- Defining IQ
- IQ Assessment
  - Assessment techniques
  - IQ aggregation and ranking
- IQ Improvement
- Wrapup

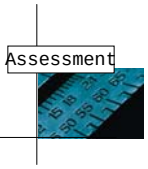


IQ Assessment



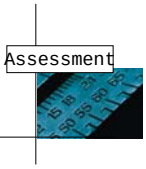
- "You cannot control what you cannot measure" [DeMarco 82]
- Why assess IQ?
  - Estimating quality, relevance, significance, ... ("garbage in/garbage out")
  - Need for improvement?
  - In case of improvement: cost-benefit ratio?

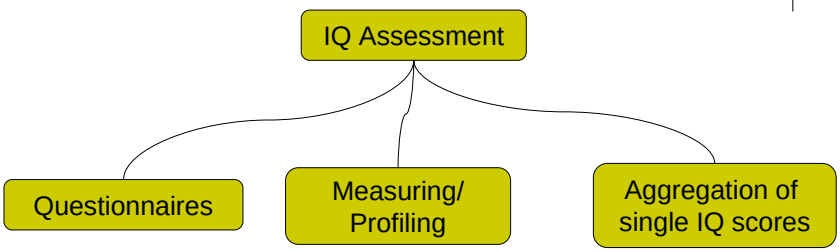
Metrics for IQ



- Measurement: quantitative comparison between an observation and a reference value
- Metrics:
  - function: IQ dimension → IQ score
- Requirements:
  - Understandable, combinable
  - Precise
  - Feasible, efficient
- But:
  - Context-specific issues → subjective measures
  - IQ values are rarely published
  - High data volume, frequent updates, ...

Assessment Techniques



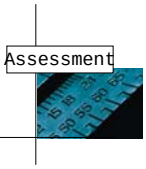


control matrices, AIMQ, ...

Objective measures, Data profiling, sampling, ...

MADM techniques

Techniques: Questionnaire



- For subjective, non-functional criteria
- Comparison to real-world state
- Exploiting human expertice
- Example: control matrices [Pierce 04]
  - Matrix for steps in data production process affecting IQ

IQ problem

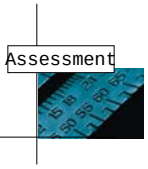
check

	duplicates	typos	missing values
Check #1	yes		
Check #2		8%	
Check #3		12%	45

rating (yes/no, category, score)

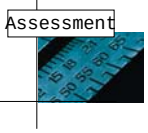
Estimating overall IQ scores by combining ratings

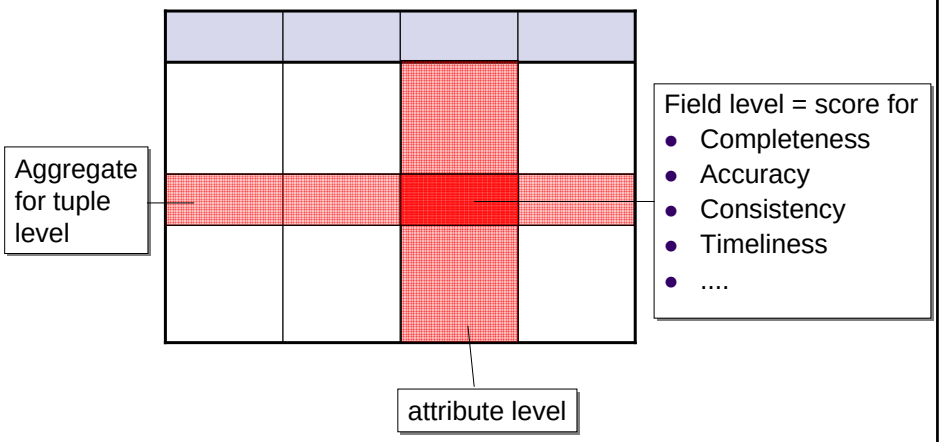
Techniques: Measuring



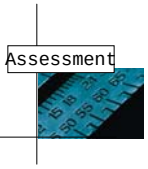
- Completeness: absence of null values, but beware of semantics of null
- Accuracy: distance between current value w and correct value w'
  - Syntactic distance: numeric values $|w-w'|$, string values edit_distance(w,w')
  - Semantic distance: Munich=München, BMW=Bayerische Motorenwerke
- Consistency: ratio of correct values wrt.
  - Integrity rules, business rules, ...
- Timeliness: 1/(update frequency • age)

Measuring



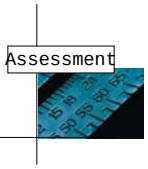


Measuring /2



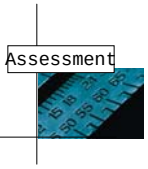
- Example: completeness of relation r with $R(A_1,A_2,...,A_n)$
  - Non-null values of A_i: $N_A = \{ t \in r \mid NotNull(t.A) \}$
  - Completeness for A_i: $\frac{|N_A|}{|r|}$
  - Completeness for $A_1,...,A_k$: $\frac{|N_{A_1,...,A_m}|}{|r|}$
  - With attribute weighting: $\frac{\sum_{t \in r} (\sum_{i=1}^{n} w_i \cdot NotNull(t(A_i)))}{|r|}$

Aggregation of Measurings



- Ratio: non-null values vs. total cardinality
  - For completeness, accuracy, ...
- Minimum/maximum
  - For timeliness, response time, ...
- Sum
  - For access costs, ...
- Product
  - For availability, ...

Combining Multiple IQ Dimensions



- IQ score = vector of (completeness, exactness, ...)
- How to compare IQ scores?

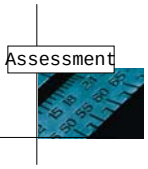
0.1,0.3,0.9,0.4,... < 0.2,0.2,0.8,0.45,... ?

IQ dimensions with different
- scales
- ranges
- importance

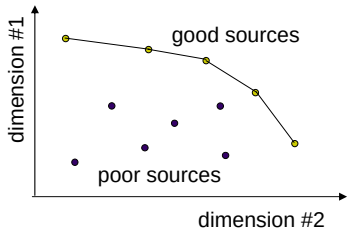
Therefore
- convert
- scale/normalize
- weight

- Multi attribute decision making (statistical techniques)
  - E.g. simple additive weighting, data envelopment analysis

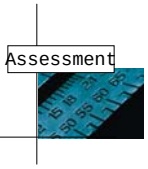
Aggregation and Ranking Methods

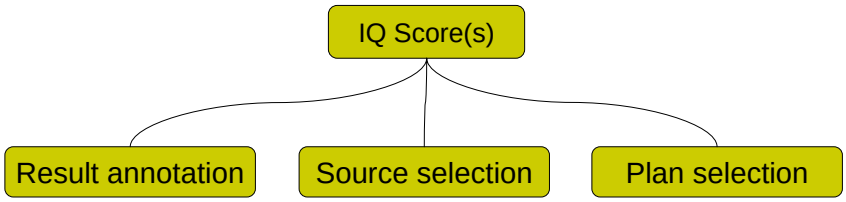


- Simple Additive Weighting
  - Scaling $v_{ij} = \frac{d - d^{jmin}}{d^{jmax} - d^{jmin}}$
  - Weighting and scoring $score = \sum_j w_j v_{ij}$ with $\sum_j w_j = 1$
- Data Envelopment Analysis
  - Does no compute a score, but suggests ranking (e.g. for data sources)



IQ Interpretation



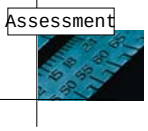


Determine quality of query result, e.g. for ranking

Query only sources fulfilling given quality constraints

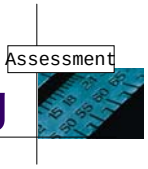
Quality as cost parameter in query optimization

Interpretation: Result Annotation



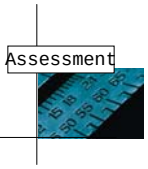
- [Motro Rakov 1998]
  - Estimating result quality for individual queries based on quality specifications (soundness, completeness)
  - Given view v:
    - Complete: if $v \supseteq v_{ideal}$
    - Sound: if $v \subseteq v_{ideal}$
  - Sampling of given and ideal view, human expertise
  - Partioning of views into homogeneous fragments (same quality value) $\rightarrow$ tree with homogeneous leafs
  - Quality estimation for simple queries (σ, π, $\times$)

Interpretation: Source Selection and Ranking



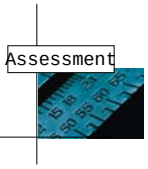
- [Mihaila et al. 2000]
  - Data sources export quality values („source content quality descriptions“)
  - Dimensions: completeness, recency, update frequency, granularity
  - Queries:
    - Data fulfilling quality requirements (e.g. completeness, granularity)
    - Source providing data with given quality (ranked on quality values)
    - Combination of sources (union)

Interpretation: Plan Selection



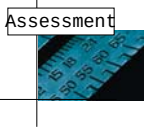
- [Naumann et al. 1999]
- Goal: choose a query plan which maximizes IQ
- IQ as additional cost parameter (beside execution costs)
- But:
  - Several IQ dimensions
  - Preferences?
- Approach
  - Prune poor sources (e.g. by applying DEA)
  - Use IQ score to rank plans (e.g. by user preferences for certain dimensions)

References: Assessment



- [Ballou Tayi 1999] D. Ballou, G. Tayi: Enhancing Data Quality in Data Warehouse Environments, Communications of the ACM 42(1): 73-78, 1999.
- [Naumann Rolker 2000] F. Naumann, C. Rolker: Assessment Methods for Information Quality Criteria, Proc. of the Int. Conference on Information Quality (IQ 2000), pages 148-162, 2000.
- [Naumann et al. 2004] F. Naumann, J. Freytag, U. Leser: Completeness of Integrated Information Sources, Information Systems 29(7):583-615, 2004.
- [Pierce 2004] E. Pierce: Assessing Data Quality with Control Matrices, Communications of the ACM 47(2): 82-86, 2004.
- [Pipino et al. 2002] L. Pipino, Y. Lee, R. Wang: Data Quality Assessment, Communications of the ACM 45(4): 211-218, 2002.

References: Interpretation



- [Chen et al. 1998] Y. Chen, Q. Zhu, N. Wang: Query Processing with Quality Control in the World Wide Web, World Wide Web Journal 1(4):241-255, 1998.
- [Mihaila et al. 2000] G. Mihaila, L. Raschid, M.-E. Vidal: Using Quality of Data Metadata for Source Selection and Ranking, Proc. of WebDB'2000, pages 93-98, 2000.
- [Naumann et al. 1999] F. Naumann, U. Leser, J. Freytag: Quality-driven Integration of Heterogeneous Information Systems, Proc. of the 25th VLDB Conference, Edinburgh, Scotland, pages 447-458, 1999.
- [Motro Rakov 1998] A. Motro, I. Rakov: Estimating the Quality of Databases, Proc. of the Int. Conference on Flexible Query Answering (FQAS 1998), pages 298-307, 1998.

Information Quality: Fundamentals, Techniques, and Use

Part 4: IQ Improvement

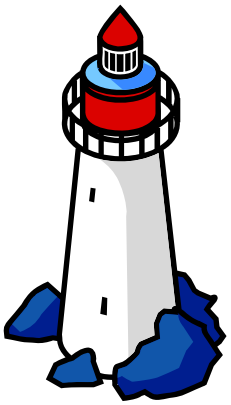
Felix Naumann
Humboldt-Universität zu Berlin

Kai-Uwe Sattler
TU Ilmenau

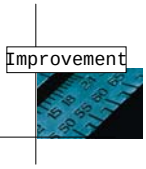
EDBT Tutorial, Munich, March 28 2006

Overview

- Motivation
- Defining IQ
- IQ Assessment
- IQ Improvement
  - Cleaning Steps
  - Profiling and Data Scrubbing
  - Outlier Detection
  - Duplicate Detection
- Wrapup

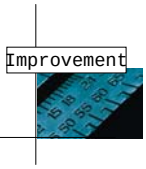


Data Cleaning



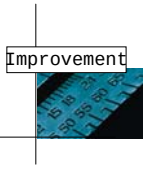
- Identifying & eliminating inconsistencies, discrepancies and errors in data in order to improve quality
- aka „data cleansing“ or „data scrubbing“
- Up to 80% of costs in DW projects
- Cleaning in data warehousing
  - As part of the ETL process
- Cleaning in information integration systems
  - „on the fly“ for virtually integrated data
  - Sometimes requires materialization

Avoiding dirty data in DBMS



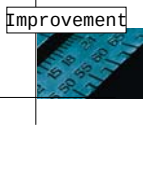
avoiding	by
wrong data types	Data type definition, `DOMAIN` constraints
wrong values	`CHECK`
missing values	`NOT NULL`
invalid references	`FOREIGN KEY`
duplicates	`UNIQUE`, `PRIMARY KEY`
inconsistencies	ACID transactions
outdated data	replication, materialized views

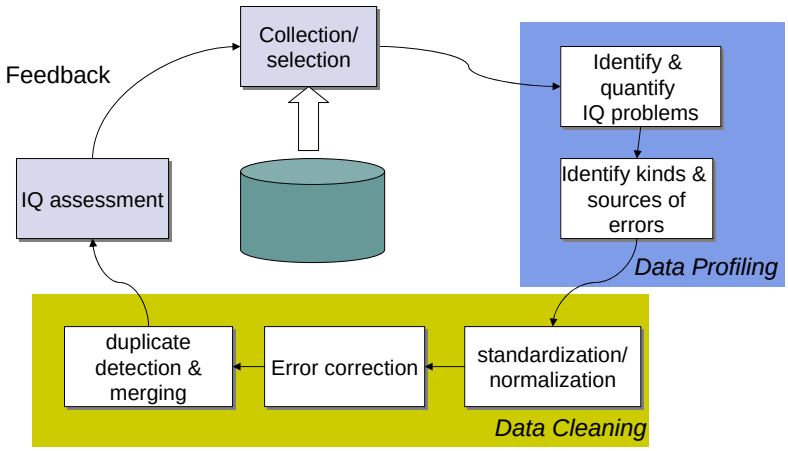
So, why is data still dirty?



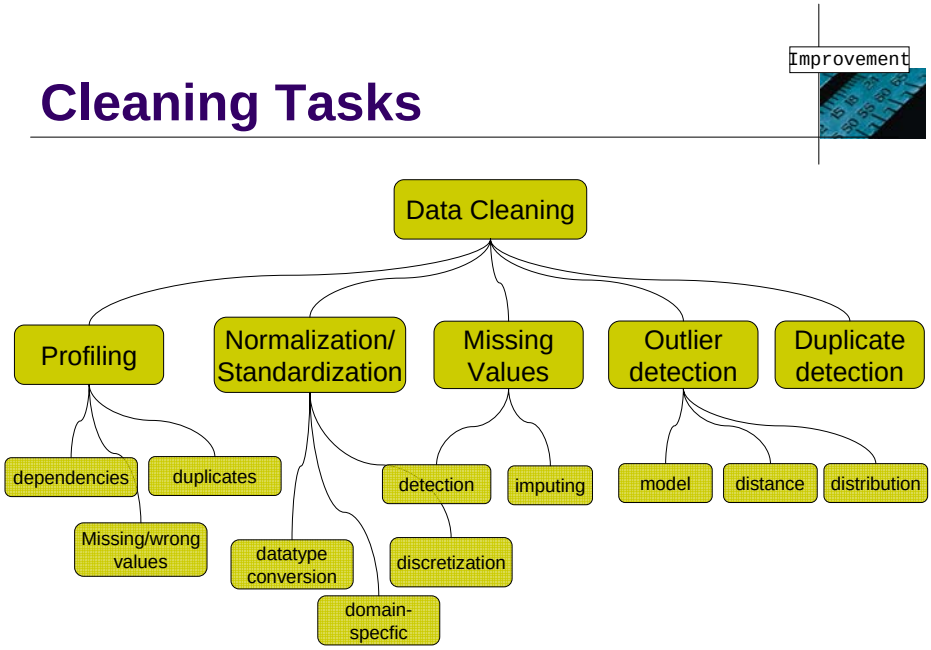
- Missing metadata, integrity constraints, ...
- Data from „foreign“ sources
- Non-DBS sources
- typos, lack of knowledge, ...
- Multi source problems, heterogeneities

Steps of Data Cleaning

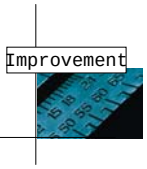




Cleaning Tasks

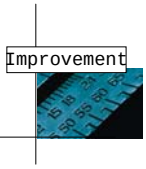


Profiling



- Analysis of content and structure of attributes
  - Data type, domain, data distribution and variance, occurence of null values, uniqueness, pattern (e.g. mm/dd/yyyy)
- Analysis of dependencies between attributes of a single relation
  - Functional dependencies, primary key candidates, „fuzzy“ dependencies
- Analysis of overlapping attributes from different relations
  - Redundancies, foreign keys

Profiling /2

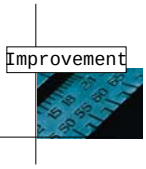


- Missing or wrong values
  - current vs.expected cardinality (e.g. number of shops, gender of customers)
  - frequency of null values, minimum / maximum, variance
- Data and input errors
  - Sorting and manual inspection
  - Similarity checks
- Duplicates
  - Number of tuples vs. Cardinality of attribute domain

Profiling: Fuzzy Dependencies

- „Fuzzy“ keys, functional dependencies and joins
  - no explicitly defined integrity constraints
  - But satisfied in most cases
- Examples
  - Primary key properties of attributes
  - Functional dependencies
  - Join paths, e.g.
    customer → profession → income class

Fuzzy Dependencies

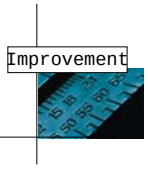


- TANE [Huhtala et al. 1999]
  - error $e(X \rightarrow A)$: minimal number of tuples which have to be eliminated to satisfy $X \rightarrow A$
  - Approach:
    1. Starting with single attributes
    2. Add addional attributes incrementally
    3. check dependencies & pruning
  - Efficient check of dependencies by partitioning the relation into equivalence classes (sets of tuples containing the same attribute values)

Profiling with SQL

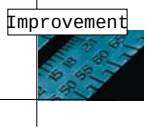
- SQL queries for basic profiling tasks
  - schema, data types: querying data dictionary
  - Domain of data
    ```
    SELECT MIN(A), MAX(A), COUNT(DISTINCT A)
    FROM DataTable
    ```
  - Erroneous data, default values
    ```
    SELECT City, COUNT(*) AS Cnt
    FROM Customer
    GROUP BY City ORDER BY Cnt
    ```
    - ascending: typos, e.g. Illmenau: 1, Ilmenau: 50
    - descending: undocumented default values, e.g. AAA: 80

Data transformation and normalization



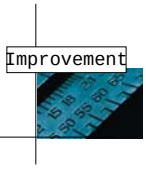
- Data type conversion: varchar → int
- Normalization: mapping into a common format
  - date: 03/01/05 → 01-MAR-2005
  - currency: $ → €
  - Uppercase strings
  - tokenizing: „Date, Chris“→ „Date“, „Chris“
- Discretization of numerical values
- Domain-specific transformations
  - Codd, Edgar Frank → Edgar Frank Codd
  - St. → Street
  - Address transformation using address databases
  - Domain-specific product names/codes (e.g. in pharmacy)

Missing Data



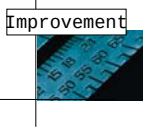
- Missing information on different levels
  - Instance level: values, tuples, relation fragments, ...
  - Schema level: Attributes, ...
- Main problems on instance level:
  - Treating null values: missing value or default value?
  - „truncation“ of values
  - Biased data, e.g. caused by null values

Detecting missing values



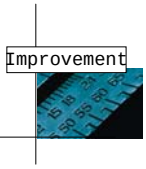
- Basic analysis:
  - Number of null values, duplicates, mean, frequency, ...
  - Comparing with expected values
    - In data warehousing on different levels of aggregation
  - Analyzing order of tuples
    - No sales information during 03/01...03/04?
    - No products with price > 20 €?

Detecting missing values /2



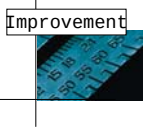
- Incomplete data, e.g. truncated and censored data
  - Sales with < 1 € are not collected in the dataset
  - Sales with > 100 € stored as 100 €
- Detection
  - By analyzing data distribution
  - But often domain knowledge required

Imputing missing values

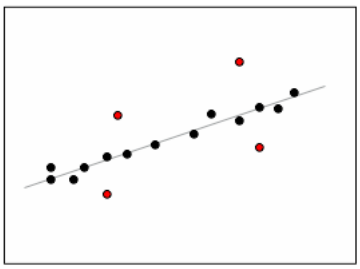


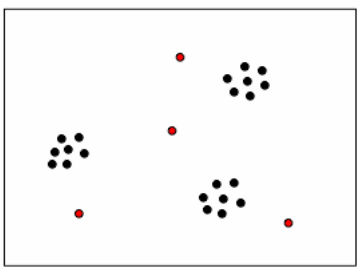
- „Unbiased estimators“
  - Estimating missing values without changing characteristics of existing dataset (mean, variance, ...)
  - E.g.: 1, 2, 3, _, 5 → (mean: 2.75; variance: 4.659)
- Exploiting functional dependencies
  - E.g.: #Bedrooms → Income
- Techniques from statistics
  - Linear regression:
    income = c • #Bedrooms
  - techniques for non-linear dependencies:
    - Neural networks, ...

Outlier detection

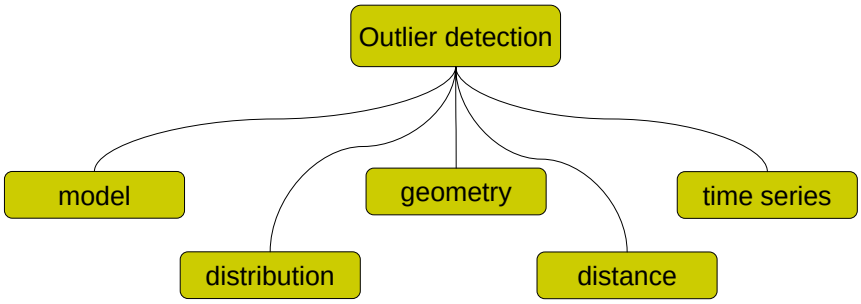


- Outlier: „suspicious“ observation that deviates too much from other observations
- Issues:
  - detection: distribution, „geometry“, time series
  - interpretation: data or observation error vs. real event





Outlier detection /2



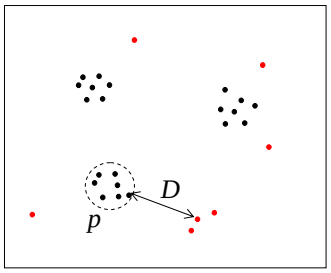
Outlier detection /3

- Model: attribute interrelationships
  - Regression
  - Rules
- Distribution / statistics
  - Based on assumption of data distribution
- Geometry
  - Points on the periphery of the dataset
  - Expensive, not applicable to higher dimensional dataset
- Distance [Knorr Ng 1998]
  - Based on distance between data points (metric distance function)
  - For higher dimension, if dataset does not fit any standard data distribution
- Time series [Dasu et al. 2000]

Distance-based outliers

Improvement

- Object o in dataset T is a $DB(p,D)$-outlier, if at least a fraction p of T lies greater than distance D from o
  [Knorr Ng 1998]
  - Outlier = object with not enough neighbors
  - Parameter p for determining „cluster of outliers“

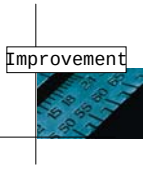


Distance-based outliers /2

Improvement

- Index-based detection (R tree, kd(B) tree)
  - Multidimensional index for determining D-neighbourhood
  - M = maximum number of objects in D-neighbourhood; if $M+1$ objects $\rightarrow$ not an outlier
  - $O(dN^2)$
- Nested loops
  - Similar, but avoid index building by smart block reading
- Cell-based
  - Partition data space into cells with two layers for each cell
  - Cell-wise check: $O(c^d+N)$

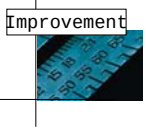
Tools for data cleaning



[Barateiro Galhardas 05]

- Auditing & Profiling
  - Axio (EvokeSoft), WizWhy (WizSoft), DB-Examiner (DBE Software), ...
- Transformation
  - SQL Server 2005, Oracle Warehouse Builder, Hummingbird ETL, ...
- Cleaning & Duplicate elimination
  - Trillium, dfPower (DataFlux), WizRule & WizSame, FirstLogic, Sagent, ...

References



- [Barateiro Galhardas 2005] J. Barateiro, H. Galhardas: A Survey of Data Quality Tools, Datenbank-Spektrum, 5(14):15-21, 2005.
- [Dasu Johnson 2003] T. Dasu, T. Johnson: Exploratory Data Mining and Data Cleaning, Wiley, 2003.
- [Dasu et al. 2000] T. Dasu, T. Johnson, E. Koutsofios: Hunting Data Glitches in Massive Time Series, Proc. of the Int. Conference on Information Qality (IQ 2000), pages 190-199, 2000.
- [Fan et al. 2001] W. Fan, H. Lu, S. Madnick, D. Cheung: Discovering and Reconciling Value Conflicts for Numerical Data Integration, Information Systems 26(8):635-656, 2001.
- [Galhardas et al. 2001] H. Galhardas, D. Florescu, D. Shasha, E. Simon, C. Saita: Declarative Data Cleaning: Language, Models and Algorithms, Proc. of the 27th VLDB Conference 2001, Roma, Italy, pages 371-380, 2001.
- [Huhtala et al. 1999] Y. Huhtala, J. Kärkkäinen, P. Porkka, H. Toivonen: TANE: An Efficient Algorithm for Discovering Functional and Approximate Dependencies, The Computer Journal 42(2):100-111, 1999.
- [Knorr Ng 1998] E. Knorr, R. Ng: Algorithms for Mining Distance-based Outliers in Large Datasets, Proc. of the 24th VLDB Conference 1998, New York, USA, pages 392-403, 1998.
- [Pyle 1999] D. Pyle: Data Preparation für Data Mining, Morgan Kaufmann Publishers, 1999.
- [Rahm Do 2000] E. Rahm, H. Do: Data Cleaning: Problems and Current Approaches, IEEE Data Engineering Bulletin, 23(4):3-13, 2000.

Information Quality: Fundamentals, Techniques, and Use
Part 4: IQ Improvement

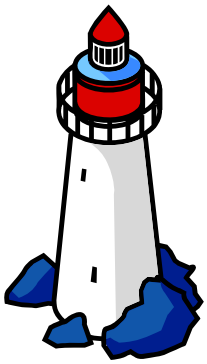
Felix Naumann
Humboldt-Universität zu Berlin

Kai-Uwe Sattler
TU Ilmenau

EDBT Tutorial, Munich, March 28 2006

Overview

- Motivation
- Defining IQ
- IQ Assessment
- IQ Improvement
  - Data Scrubbing
  - Outlier Detection
  - Duplicate Detection
- Wrapup



Duplicate Detection

First name	Last name	Address	ID
Sal	Stolpho	123 First St.	456780
Mauricio	Hernandez	321 Second Ave	123456
Klemens	Böhm	Hauptstr. 11	987654
Sal	Stolfo	123 First Street	456789

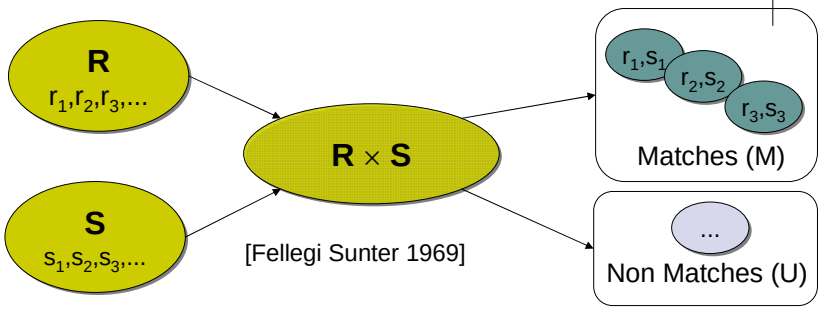
Motivation

- Possible effects
  - Example: Portfolio Management Offers
  - Credit maximum not detected
  - Too low inventory levels
  - No quantity discount for multiple orders
  - Total revenue of preferred customers unknown
  - Multiple mailings of same catalog to same household
- General problems
  - Additional, unnecessary IT expenses
  - Low customer satisfaction
  - Potentials and dangers not detected
  - Poor quality financial data

“Duplicate Detection” has many Duplicates

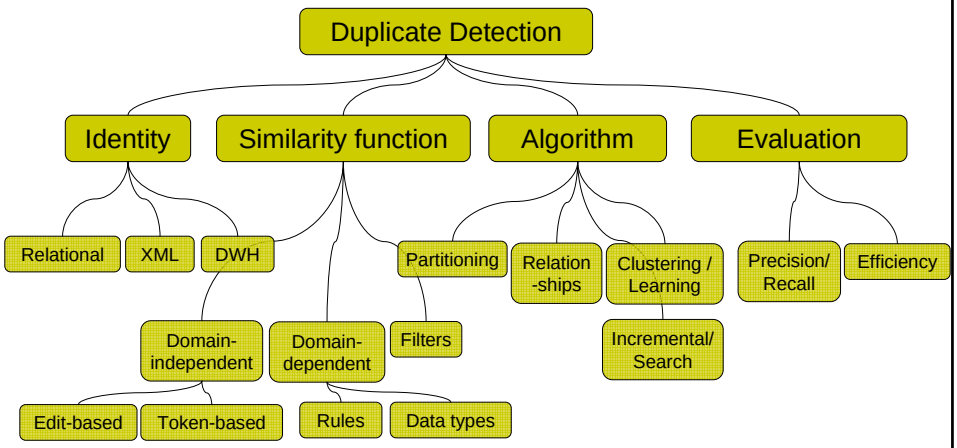
- Duplicate detection / de-duplication
- Record linkage
- Object identification / object consolidation
- Entity resolution / entity clustering
- Reference reconciliation / reference matching
- Householding / household matching
- Match / Fuzzy match / approximate match
- Merge/purge
- Hardening soft databases
- Identity uncertainty
- “mixed and split citation problem”

The Problem



- Duplicate Detection (Record Linkage)
  - Identification of semantically equivalent representations, i.e., representations of the same real-world object
- Duplicate Elimination (Reconciliation / Fusion)
  - Create a complete, concise, and consistent data set.

Duplicate Detection



Similarity functions – domain independent

- Edit-based
  - Edit-distance / Levenshtein-distance [Levenshtein 1965]
    - Minimum number of edits from one word to the other
    - Domain-specific costing
    - Also: Smith-Waterman [Smith Waterman 1981]
      - Compensates abbreviations
  - Soundex
    - 4-letter code for each word
    - SOUNDEX('Farwick ') = F620
      - Fahruschi, Feuerhake, Frass, Fricke
  - Jaro [Jaro 1989] / Jaro-Winkler [Winkler 1999]
    - Common letters within ½ string length
    - Transposed letters
  - SQL LIKE
    - Precision / Recall tradeoff
      - Fr% vs. Frick%
    - Expensive, no similarity scoring
- Token-based
  - Tokens
    - Words / Terms
    - n-grams
  - Jaccard
    - |{common tokens}| / |{all tokens}|
  - TFIDF [Cohen et al. 2003]
    - Term frequency (tf)
    - Inverse document frequency (idf)
    - Tfidf: log (tf+1) x log idf
    - Common words have low weight
    - Cosine similarity of term vectors weighted by tfidf
  - And many more [Koudas Srivastavasa 2005]

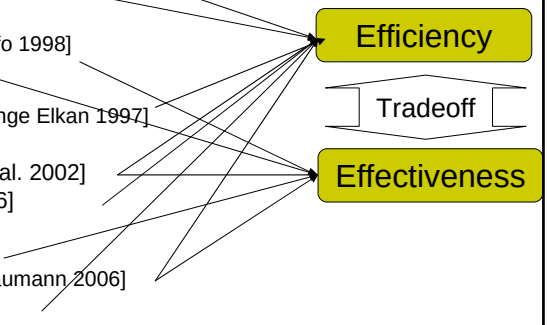
Similarity functions – domain dependent

- Data Types
  - Special similarity for dates
  - Special similarity for numerical attributes
  - ...
- Rules
  - [Hernandez Stolfo 1998], [Lee et al. 2000]
  - Given two records, **`r1`** and **`r2`**.

```
IF  last name of r1 = last name of r2,
 AND first names differ slightly,
 AND address of r1 = address of r2
 THEN r1 is equivalent to r2.
```

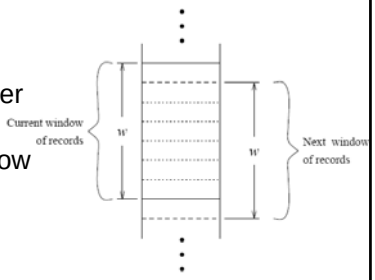

Algorithms

- Partitioning
  - Simple Partitioning / Blocking
  - Sorted Neighborhood SNM [Hernandez Stolfo 1998]
  - Extensions to SNM
    - Multipass [Hernandez Stolfo 1998]
    - Domain-independent [Monge Elkan 1997]
    - Prime representatives [Monge Elkan 1997]
- Relationships
  - DELPHI [Ananthakrishna et al. 2002]
  - SNMX [Puhlmann et al. 2006]
  - Relationship-aware
    - SEMEX [Dong et al. 2005]
    - ReconA / AdamA [Weis Naumann 2006]
- Incremental / Search



Sorted Neighborhood

- Idea
  - Sort tuples so that similar tuples are close to each other.
  - Only compare tuples within a small neighborhood (window)
- Generate key
  - E.g.: SSN+“first 3 letters of name“ + ...
- Sort by key
  - Similar tuples end up close to each other
- Slide window over sorted tuples
  - Compare all pairs of tuples within window
- Problems
  - Choice of Key
  - Choice of window size
- Complexity: at least 3 passes over data
  - Sorting!



[Hernandez Stolfo 1998]

Sorted Neighborhood Extensions

- Multi-Pass [Hernandez Stolfo 1998]
  - Several runs with different keys
  - Smaller window
  - Transitive closure over different runs
- Domain-independent [Monge Elkan 1997]
  - 1st pass: Key = tuple
  - 2nd pass: Key = reversed tuple
  - Similarity: Smith-Waterman
- Prime representatives [Monge Elkan 1997]
  - Clusters of duplicates
  - Compare new tuple only with some prime representative of a cluster

DELPHI – Data Warehouse Duplicates

ID	country
1	USA
2	United States
3	Unitd States

ID	City	Country_ID
1	New York	1
2	Los Angeles	1
3	Now York	2
4	Los Angeles	2
5	New York	3
6	Los Angels	3

String similarity
→ 2 ≈ 3

Common children
→ 1 ≈ 2 ≈ 3

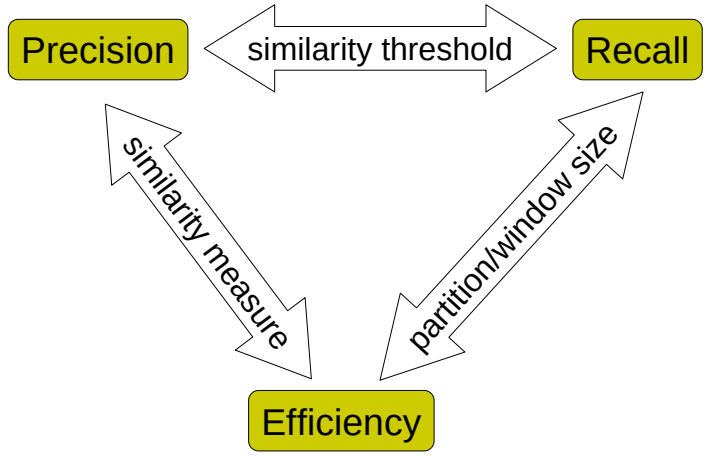
[Ananthakrishna et al. 2002]

SNMX – Sorted Neighborhood for XML

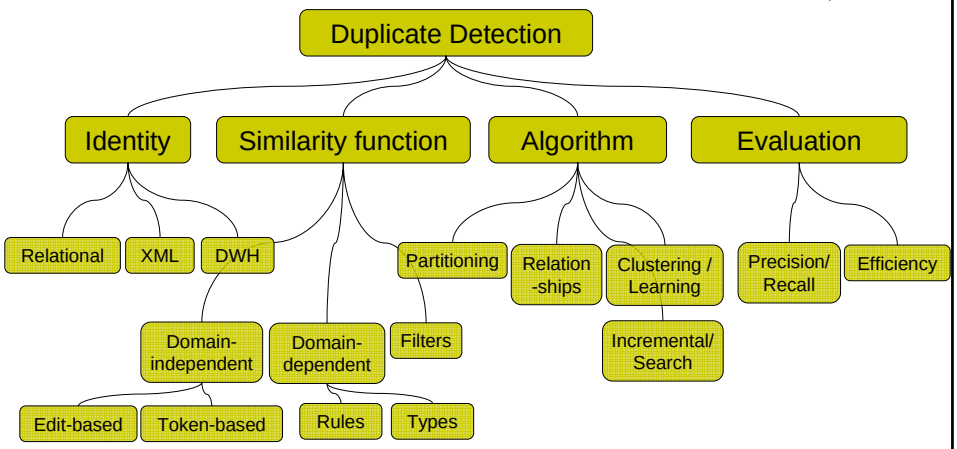
- Delphi: Top-down
- SNMX: Bottom up
  - Idea: Apply SNM to each hierarchy level beginning at bottom
  - Intuition: Only elements with many duplicate children need to be compared
  - Increased efficiency

[Puhlmann et al. 2006]

Evaluating Duplicate Detection



Duplicate Detection – Summary



Data Fusion / Reconciliation

- Duplicate elimination
  - Keep any tuple
  - Keep best tuple
    - Subsumption
    - Highest quality tuple
- Duplicate fusion
  - Conflicts among duplicates
  - Conflict resolution functions
- XML Data fusion

References

- [Ananthakrishna et al. 2002]
  Rohit Ananthakrishna, Surajit Chaudhuri, Venkatesh Ganti: Eliminating Fuzzy Duplicates in Data Warehouses. VLDB 2002: 586-597
- [Cohen et al. 2003]
  William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg. A comparison of string distance metrics for name-matching tasks. In Proceedings of the IJCAI Workshop on Information Integration on the Web (IIWeb), pages 73{78, 2003.
- [Dong et al. 2005]
  Xin Dong, Alon Y. Halevy, Jayant Madhavan: Reference Reconciliation in Complex Information Spaces. SIGMOD Conference 2005: 85-96
- [Fellegi Sunter 1969]
  I. Fellegi, A. Sunter: A theory of record linkage. Journal of the American Statistical Association, Vol 64. No 328, 1969
- [Hernandez Stolfo 1998]
  M. Hernandez, S. Stolfo: Real-world Data is Dirty: Data Cleansing and the Merge/Purge, Journal of Data Mining and Knowledge Discovery, 2(1):9-37, 1998.
- [Jaro 1989]
  M. A. Jaro: Advances in record linkage methodology as applied to matching the 1985 census of Tampa, Florida. Journal of the American Statistical Association 84: 414-420.
- [Koudas Srivastavasa 2005]
  Nick Koudas and Divesh Srivastavasa: Approximate Joins: Concepts and Techniques. Tutorial at VLDB 2005
- [Lee et al. 2000]
  M. Lee, T. Ling, W. Low: IntelliClean: A Knowledge-based Intelligent Data Cleaner, Proc. ACM SIGKDD Conference 2000, Boston, USA, pages 290-294, 2000.

References

- [Levenshtein 1965]
  Vladimir Levenshtein: Binary codes capable of correcting spurious insertions and deletions of ones. Problems of Information Transmission vol 1, 1965, 8-17
- [Monge Elkan 1997]
  Alvaro Monge, Charles Elkan: An Efficient Domain-Independent Algorithm for Detecting Approximately Duplicate Database Records, In Porceedings of the Workshop on Research Issues on Data Mining and Knowledge Discovery, Tucson, AZ, 1997
- [Puhlmann et al. 2006]
  Sven Puhlmann, Melanie Weis, Felix Naumann: XML Duplicate Detection Using Sorted Neigborhoods, Proceedings of the International Conference on Extending Database Technology (EDBT) 2006, Munich, Germany
- [Smith Waterman 1981]
  T.F. Smith and M.S. Waterman: Identification of common molecular subsequences. Journal of Molecular Biology, Vol. 147, 1981
- [Weis Naumann 2006]
  Melanie Weis and Felix Naumann: Detecting Duplicates in Complex XML Data, in Proceedings of the International COnference in Data Engineering 2006, Atlanta, GA. Poster
- [Winkler 1999]
  William E. Winkler: The state of record linkage and current research problems. IRS publication R99/04 (http://www.census.gov/srd/www/byname.html)

Information Quality: Fundamentals, Techniques, and Use

Part 5: Wrap Up

Felix Naumann
Humboldt-Universität zu Berlin

Kai-Uwe Sattler
TU Ilmenau

EDBT Tutorial, Munich, March 28 2006

Motivation

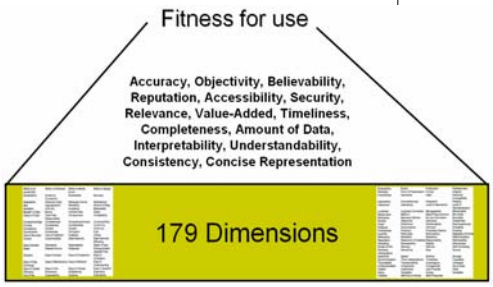
- Poor information quality costs money (and more).
- Poor information quality is a fact.
- Thus:
  - Define IQ
  - Assess IQ
  - Improve IQ

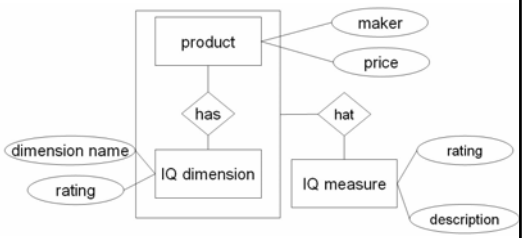
Goodyear reveals $100 million error 10/23/07 USA

Goodyear said late Wednesday that it will restate earnings for the past five years, decreasing income by as much as $100 million because an accounting system caused billing errors. The tiremaker is delaying the release of its third-quarter earnings, expected this morning, until mid-November. Shares closed up 2 cents to $6.83 before the announcement; in after-hours trading, shares plummeted 27%, or $1.83, to $5.

Defining IQ

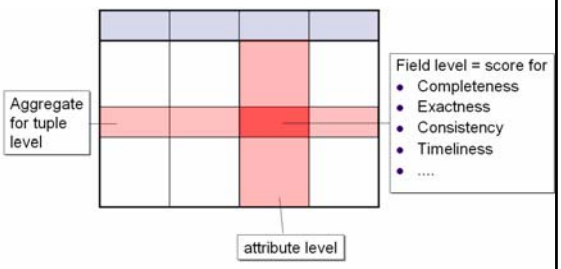
- IQ dimensions
- IQ classifications
- IQ in data models





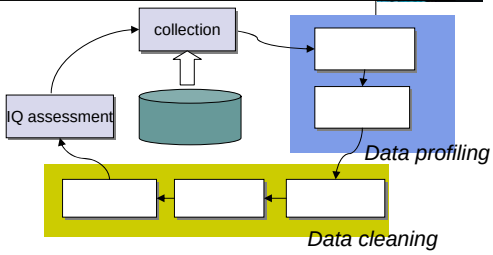
IQ Assessment

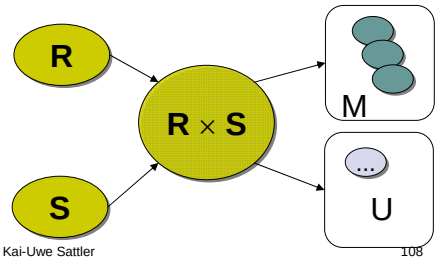
- Techniques
  - Questionnaires
  - Measuring / Profiling
  - Aggregation
- IQ interpretation
  - Annotation
  - Source selection
  - Plan selection



IQ Improvement

- Data scrubbing
- Outlier detection
- Duplicate detection





IQ Community

- Conferences and workshops
  - ICIQ (@ MIT, Boston)
    - 11th International Conference on Information Quality
    - Deadline: July 8
  - IQIS (@ SIGMOD)
    - 3rd SIGMOD Workshop on Information Quality in Information Systems
    - Deadline April 14
  - CleanDB (@ VLDB): Deadline June 2
  - Others
    - QoIS 2006 (@ ER) Quality of Information Systems
    - DIQ (@ CAiSE) Workshop on Data and Information Quality (DIQ)
    - WISQ (@ WISE) Web Information Systems Quality Workshop
- Organisations
  - MITs TDQM: http://web.mit.edu/tdqm/www/
    - ICIQ conference, workshops, courses
    - Master of Science in Information Quality (MSIQ) at University of Arkansas at Little Rock
  - Deutsche Gesellschaft für Informations- und Datenqualität e.V. www.dgiq.de
    - Deutsche Information Quality Management Konferenz & Workshop
    - German IQ Community
    - IQM-Contest

Areas of Interest

- Database-related
  - IQ assessment
  - Duplicate detection, data cleansing and reconciliation
  - Customer data integration, householding
  - Data integration and fusion
  - Data quality and cleaning in information extraction, semi-structured data, multimedia data, graphs, and sensor data
  - Quality-aware aware query languages and query processing
  - Detection of contradictory data, outliers, inconsistencies, noise
  - Mining for patterns of poor quality data
  - IQ in scientific data management
  - Application-driven Information Quality: Bioinformatics, Marketing, CRM, e-Business, Geomedia, etc.

Further areas of Interest

- Conceptual
  - IQ Concepts, Tools, Metrics, Measures, Models, and Methodologies
  - Information Product Implementation, Delivery, and Management
  - IQ in Databases, the Web, and e-Business
  - Trust, Knowledge, and Society in the IQ Context
  - IQ Policies and Standards
- Other
  - IQ Practices: Case Studies and Experience Reports
  - IQ Product Experience Reports
  - IQ Education and Curriculum Development
  - Economic aspects of information quality

The End

- Felix Naumann
  - Humboldt-Universität zu Berlin
  - http://www.informatik.hu-berlin.de/mac/
  - naumann@informatik.hu-berlin.de
- Kai-Uwe Sattler
  - TU Ilmenau
  - http://www.tu-ilmenau.de/dbis/
  - k.sattler@computer.org