Innovationslabor
Semantische Integration von Webdaten

Workflow-basierte Daten-integration und Objekt-Matching

Dr. Andreas Thor

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

http://dbs.uni-leipzig.de/format

Workflow-basierte Datenintegration

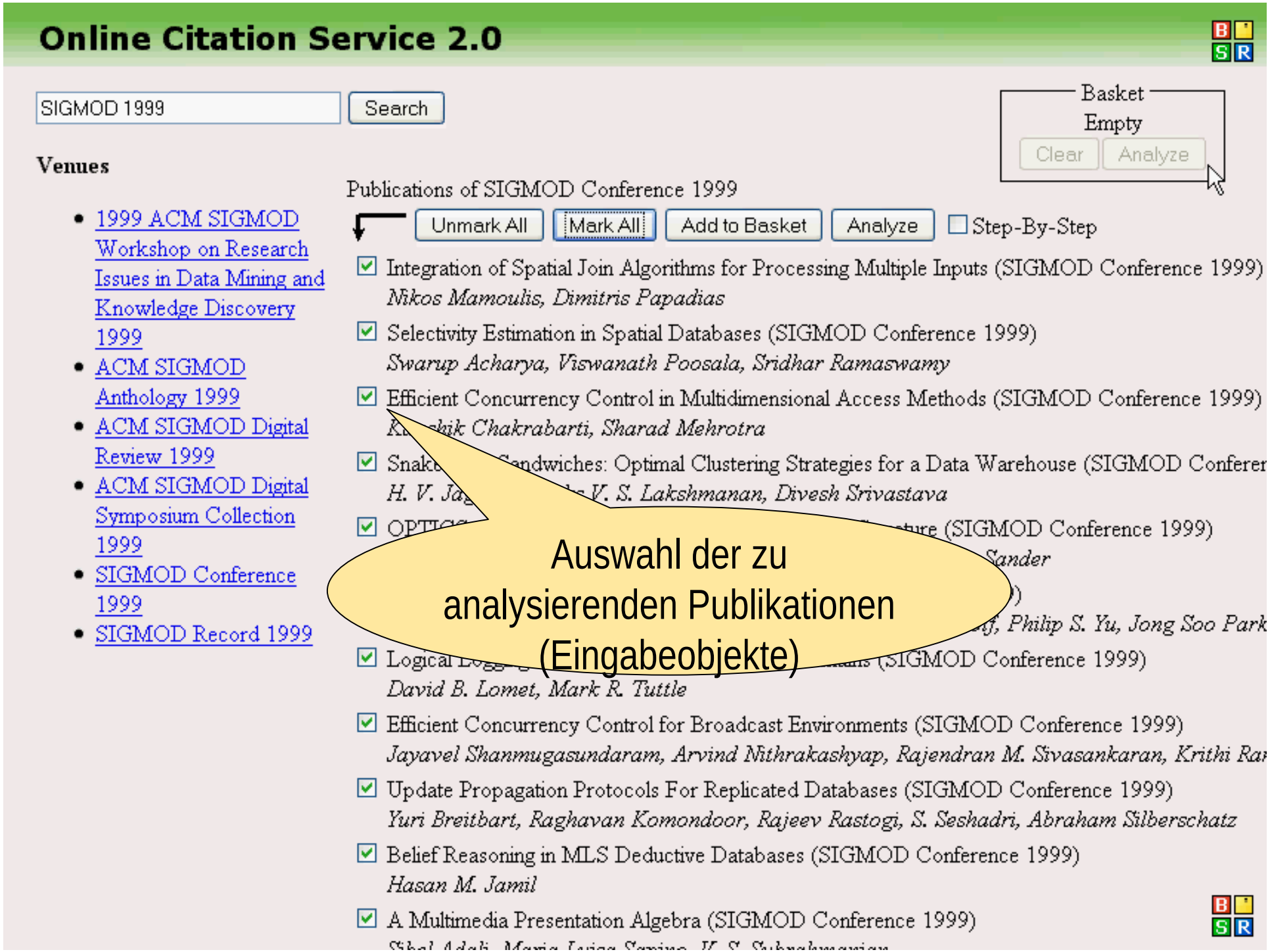
- Ausgangspunkt / Problemstellung
  - Lösung eines konkreten Integrationsproblems erfordert eine koordinierte Ausführung mehrerer Teilschritte, u.a.
    - Anfragen an Suchmaschinen / Datenquellen
    - Abgleich der Daten (Objekt-Matching) und Metadaten (Schema-Matching)
  - bisherige Werkzeuge ungeeignet
    - Programmiersprachen (Java, C#, ...) sehr “low level” → viel Code
    - ETL-Werkzeuge zu statisch / unflexibel
- Ansatz: Erweiterung des Mashup-Ansatzes zur schnellen Realisierung von Datenintegrationsaufgaben
- Forschungsarbeiten seit 2005
  - Prototyp iFuice und Beispiel-Mashup OCS

Prototyp iFuice: Features

- Workflow-artiger (programmatischer) Integrationsansatz
  → zur Lösung komplexer Integrationsaufgaben
- “Programmiersprache zur Datenintegration”
  → Operatoren für typische Aufgaben (query, match, ...)
- Nutzung bestehender Verknüpfungen (Mappings)
  → Wiederverwendung z.B. von Web-Links
- Verwendet ein flexibles Domänenmodell
  → typisierte Objekte mit Attribut-Wert-Paaren
- P2P-artige Kopplung von Datenquellen, kein zentrales Schema
  → Einfaches „Ankoppeln“ neuer Datenquellen „wo es am besten passt“ bzw. „am einfachsten ist“

Beispiel-Mashup: OCS

- Online Citation Service
  - Finden von Zitierungszahlen wissenschaftlicher Publikationen
  - Hochqualitative Zitierungsanalyse für Autoren, Konferenzen, ...
- Problemstellung / typisches Workflow-Muster
  - Finden intelligenter Suchanfragen an unterschiedliche Datenquellen, um eine Menge von Objekten (hier: Publikationen) zu finden
  - Extraktion der benötigten Informationen (hier: Zitierungszahl) aus den Ergebnissen
  - Zuordnung der gefundenen Objekte zu den Eingabeobjekten (Objekt-Matching)
- Analoge Anwendungsszenarien
  - Finden der Preise und/oder Bewertungen zu einer Menge von Produkten in anderen Datenquellen
  - Urlaubsbuchung mit abhängigen Komponenten
    Flüge → “dazu passende” Hotels → “dazu passender” Mietwagen


Online Citation Service 2.0
SIGMOD 1999
Search
Basket
Empty
Clear
Analyze
Venues
1999 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery 1999
ACM SIGMOD Anthology 1999
ACM SIGMOD Digital Review 1999
ACM SIGMOD Digital Symposium Collection 1999
SIGMOD Conference 1999
SIGMOD Record 1999
Publications of SIGMOD Conference 1999
Unmark All
Mark All
Add to Basket
Analyze
Step-By-Step
Integration of Spatial Join Algorithms for Processing Multiple Inputs (SIGMOD Conference 1999)
Nikos Mamoulis, Dimitris Papadias
Selectivity Estimation in Spatial Databases (SIGMOD Conference 1999)
Swarup Acharya, Viswanath Poosala, Sridhar Ramaswamy
Efficient Concurrency Control in Multidimensional Access Methods (SIGMOD Conference 1999)
Auswahl der zu analysierenden Publikationen (Eingabeobjekte)
Logical
David B. Lomet, Mark R. Tuttle
Efficient Concurrency Control for Broadcast Environments (SIGMOD Conference 1999)
Jayavel Shanmugasundaram, Arvind Nithrakashyap, Rajendran M. Sivasankaran, Krithi Ra
Update Propagation Protocols For Replicated Databases (SIGMOD Conference 1999)
Yuri Breitbart, Raghavan Komondoor, Rajeev Rastogi, S. Seshadri, Abraham Silberschatz
Belief Reasoning in MLS Deductive Databases (SIGMOD Conference 1999)
Hasan M. Jamil
A Multimedia Presentation Algebra (SIGMOD Conference 1999)

Online Citation Service 2.0

	All	DBLP	GS	ACM	CS	Libra
# Publications	85	85	72	84	25	90
∑ Citations			3522	1516	940	2336
∅ Citations			41.4	17.8	11.1	27.5
H-Index			23	23	17	30

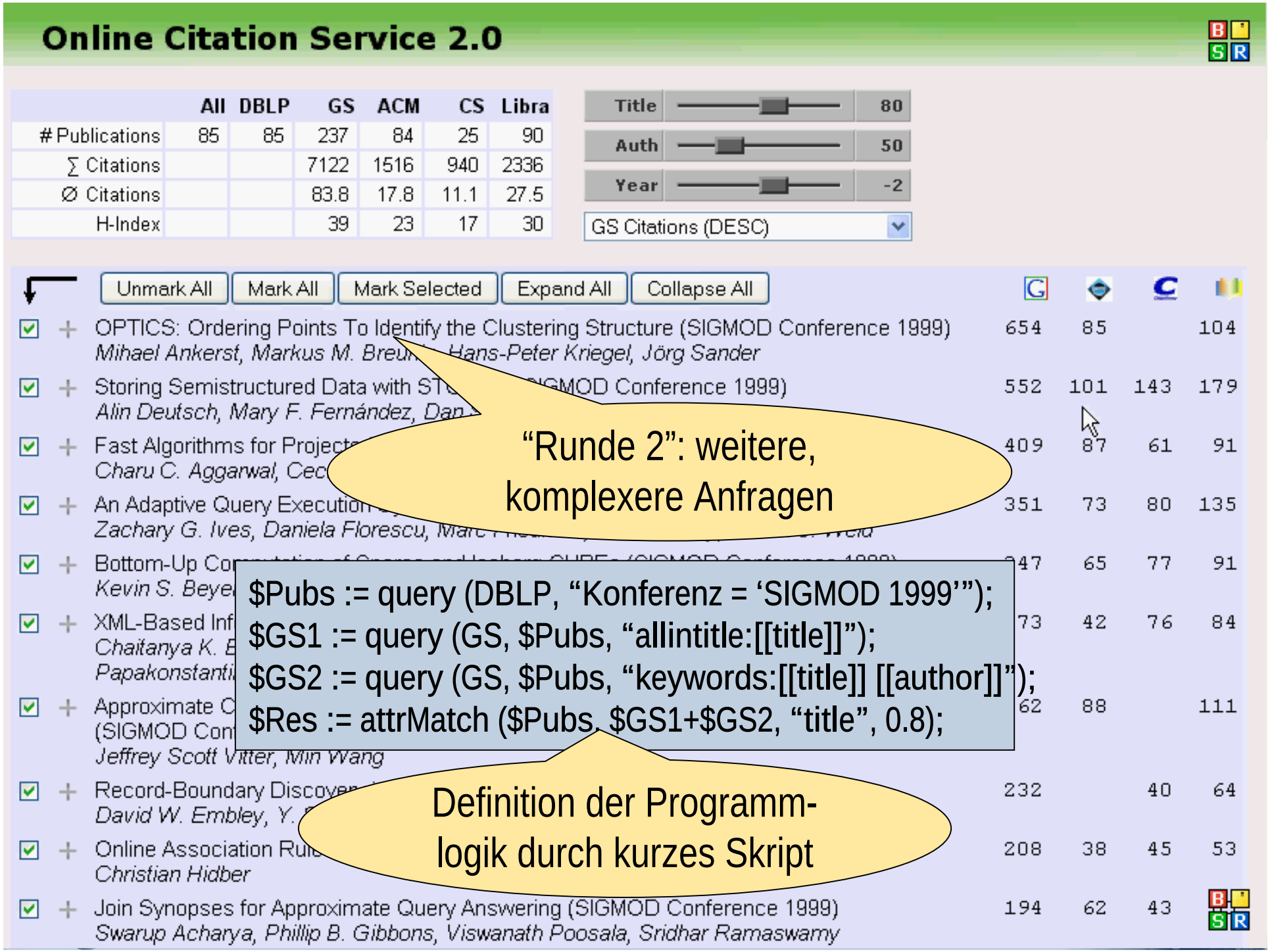
Title 80 | Auth 50 | Year -2

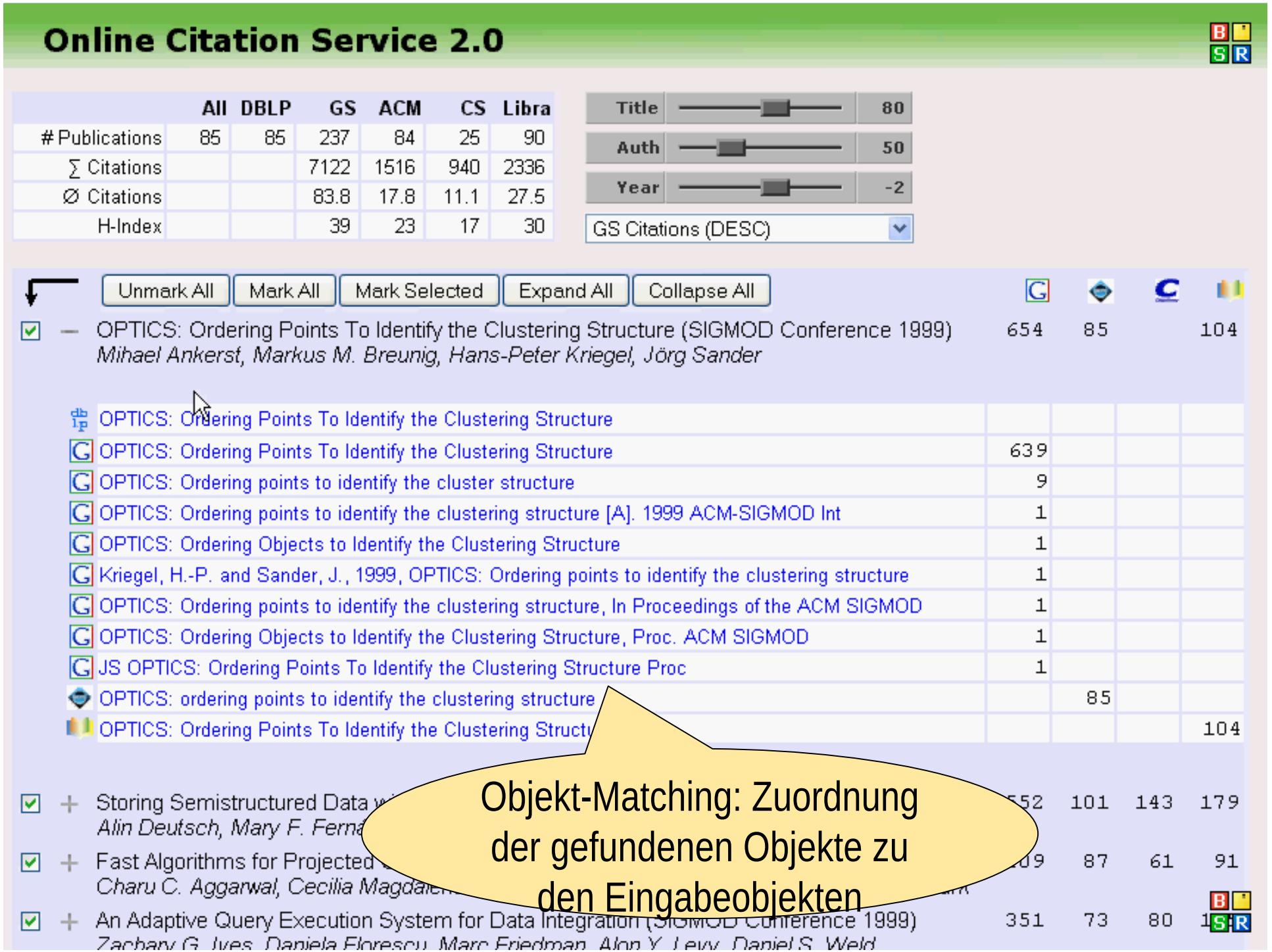
GS Citations (DESC)

Unmark All | Mark All | Mark Selected | Expand All | Collapse All

Storing Semistructured Data with STORED (SIGMOD Conference 1999) *Alin Deutsch, Mary F. Fernández, Dan Suciu*	527	101	143	179
Bottom-Up Computation of Sparse and Iceberg CUBEs (SIGMOD Conference 1999) *Kevin S. Beyer, Raghu Ramakrishnan*	344	65	77	91
An Adaptive Query Execution Syste… Data Integration (SIGMOD Conference 1999) *Zachary G. Ives, Daniela Florescu, M…man, Alon Y. Levy, Daniel S. Weld*	301	73	80	135
Record-Boundary Discovery in Web Doc… 1999) *David W. Embley, Y. S. Jiang, …*	222		40	64
Online Association Rule … *Christian Hidber*	203	38	45	53
Self-tuning Histograms: Building Histog… …nference 1999) *Ashraf Aboulnaga, Surajit Chaudhuri*	180	50		62
BOAT-Optimistic Decision Tree Construction (SIGMOD Conference 1999) *Johannes Gehrke, Venkatesh Ganti, Raghu Ramakrishnan, Wei-Yin Loh*	177	44		55
Selectivity Estimation in Spatial Databases (SIGMOD Conference 1999) *Swarup Acharya, Viswanath Poosala, Sridhar Ramaswamy*	160	47		66
On Random Sampling over Joins (SIGMOD Conference 1999) *Surajit Chaudhuri, Rajeev Motwani, Vivek R. Narasayya*	143	54		52
Query Optimization in the Presence of Limited Access Patterns (SIGMOD Conference 1999) *Daniela Florescu, Alon Y. Levy, Ioana Manolescu, Dan Suciu*	128	27	60	62
Query Rewriting for Semistructured Data (SIGMOD Conference 1999)	118	33		45

"Runde 1": wenige, einfache Anfragen


Online Citation Service 2.0
All DBLP GS ACM CS Libra
Publications 85 85 237 84 25 90
∑ Citations 7122 1516 940 2336
Ø Citations 83.8 17.8 11.1 27.5
H-Index 39 23 17 30
Title 80
Auth 50
Year -2
GS Citations (DESC)
Unmark All
Mark All
Mark Selected
Expand All
Collapse All
OPTICS: Ordering Points To Identify the Clustering Structure (SIGMOD Conference 1999) 654 85 104
Mihael Ankerst, Markus M. Breu Hans-Peter Kriegel, Jörg Sander
Storing Semistructured Data with ST GMOD Conference 1999) 552 101 143 179
Alin Deutsch, Mary F. Fernández, Dan
Fast Algorithms for Projecte 409 87 61 91
Charu C. Aggarwal, Cec
An Adaptive Query Execution 351 73 80 135
Zachary G. Ives, Daniela Florescu,
Bottom-Up Co 47 65 77 91
Kevin S. Beye
XML-Based Inf 73 42 76 84
Chaitanya K. B
Papakonstanti
Approximate C 62 88 111
(SIGMOD Con
Jeffrey Scott Vitter, Min Wang
Record-Boundary Discover 232 40 64
David W. Embley, Y.
Online Association Rul 208 38 45 53
Christian Hidber
Join Synopses for Approximate Query Answering (SIGMOD Conference 1999) 194 62 43
Swarup Acharya, Phillip B. Gibbons, Viswanath Poosala, Sridhar Ramaswamy
“Runde 2”: weitere, komplexere Anfragen
$Pubs := query (DBLP, “Konferenz = ‘SIGMOD 1999’”);
$GS1 := query (GS, $Pubs, “allintitle:[[title]]”);
$GS2 := query (GS, $Pubs, “keywords:[[title]] [[author]]”);
$Res := attrMatch ($Pubs, $GS1+$GS2, “title”, 0.8);
Definition der Programm-logik durch kurzes Skript


Online Citation Service 2.0
All DBLP GS ACM CS Libra
Publications 85 85 237 84 25 90
∑ Citations 7122 1516 940 2336
Ø Citations 83.8 17.8 11.1 27.5
H-Index 39 23 17 30
Title 80
Auth 50
Year -2
GS Citations (DESC)
Unmark All
Mark All
Mark Selected
Expand All
Collapse All
OPTICS: Ordering Points To Identify the Clustering Structure (SIGMOD Conference 1999) 654 85 104
Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, Jörg Sander
OPTICS: Ordering Points To Identify the Clustering Structure
OPTICS: Ordering Points To Identify the Clustering Structure 639
OPTICS: Ordering points to identify the cluster structure 9
OPTICS: Ordering points to identify the clustering structure [A]. 1999 ACM-SIGMOD Int 1
OPTICS: Ordering Objects to Identify the Clustering Structure 1
Kriegel, H.-P. and Sander, J., 1999, OPTICS: Ordering points to identify the clustering structure 1
OPTICS: Ordering points to identify the clustering structure, In Proceedings of the ACM SIGMOD 1
OPTICS: Ordering Objects to Identify the Clustering Structure, Proc. ACM SIGMOD 1
JS OPTICS: Ordering Points To Identify the Clustering Structure Proc 1
OPTICS: ordering points to identify the clustering structure 85
OPTICS: Ordering Points To Identify the Clustering Structu 104
Objekt-Matching: Zuordnung der gefundenen Objekte zu den Eingabeobjekten
Storing Semistructured Data
Alin Deutsch, Mary F. Ferna
552 101 143 179
Fast Algorithms for Projected
Charu C. Aggarwal, Cecilia Magdalen
109 87 61 91
An Adaptive Query Execution System for Data Integration (SIGMOD Conference 1999) 351 73 80
Zachary G. Ives, Daniela Florescu, Marc Friedman, Alon Y. Levy, Daniel S. Weld

Online Citation Service 2.0

	All	DBLP	GS	ACM	CS	Libra
# Publications	85	85	109	71	37	83
∑ Citations			6058	1365	1146	2334
Ø Citations			71.3	16.1	13.5	27.5
H-Index			35	22	19	30

Title 100
Auth 50
Year -2
GS Citatio (DESC)

Einstellung des Objekt-Matchings durch Schwellwerte

Unmark All | Mark All | Mark Selected | Expand All | e All

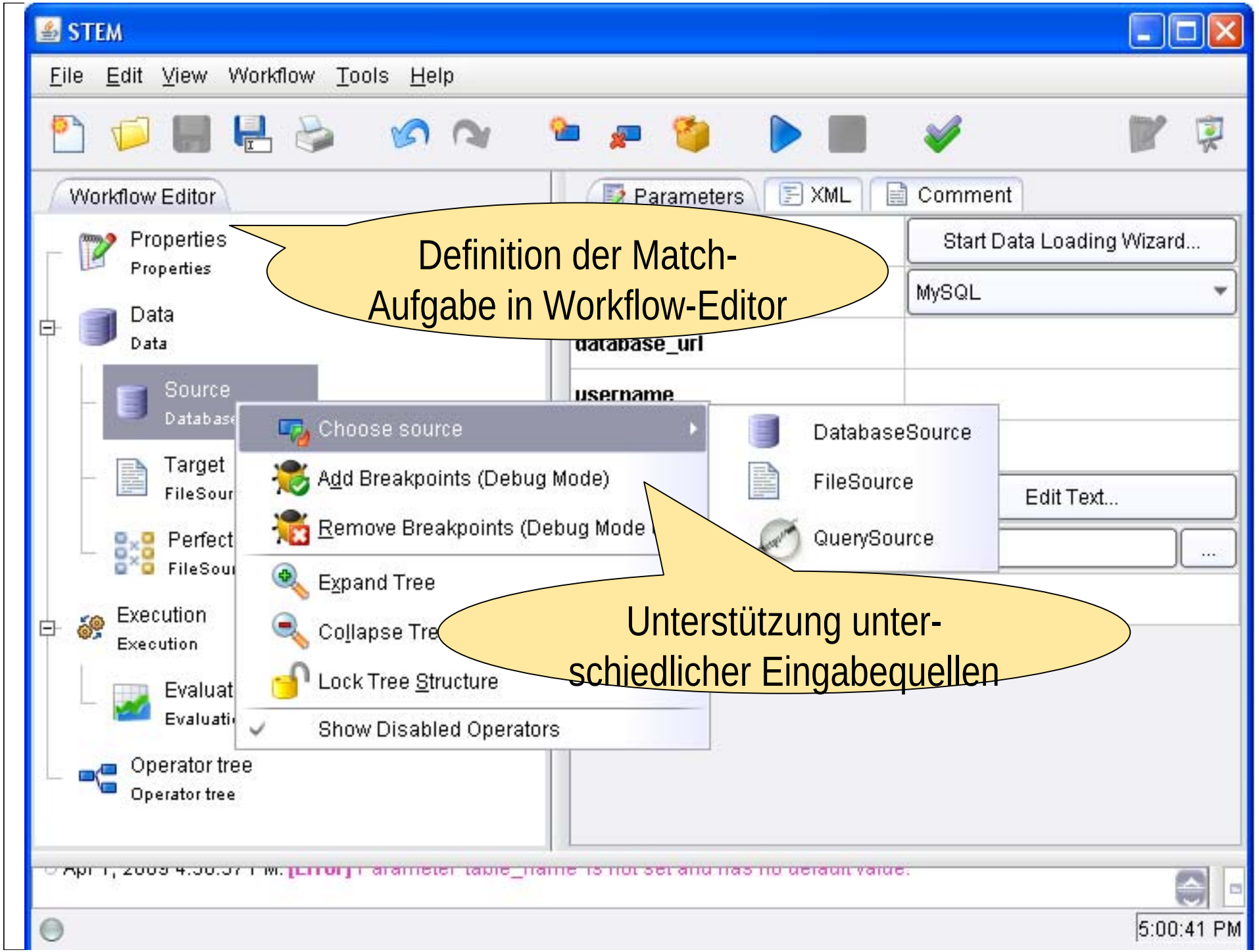
OPTICS: Ordering Points To Identify the Clustering S MOD Conference 1999) *Mihael Ankerst, Markus M. Breuni*	648	85	76	104
Storing Semistructur *Alin Deutsch, Mary*	527	101	143	179
Fast Algorithms for Projecte *Charu C. Aggarwal, Cecilia Magdalena Procopiuc, Joel L. Wolf, Philip S. Yu, Jong Soo Park*	404	87	61	91
An Adaptive Query Execution System for Data Integration (SIGMOD Conference 1999) *Zachary G. Ives, Daniela Florescu, Marc Friedman, Alon Y. Levy, Daniel S. Weld*	301	73	80	135
Approximate Computation of Multidimensional Aggregates of Sparse Data Using Wavelets (SIGMOD Conference 1999) *Jeffrey Scott Vitter, Min Wang*	259	88		111
XML-Based Information Mediation with MIX (SIGMOD Conference 1999) *Chaitanya K. Baru, Amarnath Gupta, Bertram Ludäscher, Richard Marciano, Yannis Papakonstantinou, Pavel Velikhov, Vincent Chu*	258	42	76	84
Record-Boundary Discovery in Web Documents (SIGMOD Conference 1999) *David W. Embley, Y. S. Jiang, Yiu-Kai Ng*	225		40	64
Online Association Rule Mining (SIGMOD Conference 1999) *Christian Hidber*	201	38	45	53
Ripple Joins for Online Aggregation (SIGMOD Conference 1999) *Peter J. Haas, Joseph M. Hellerstein*	188	55	57	80
Join Synopses for Approximate Query Answering (SIGMOD Conference 1999) *Swarup Acharya, Phillip B. Gibbons, Viswanath Poosala, Sridhar Ramaswamy*	186	62	43	

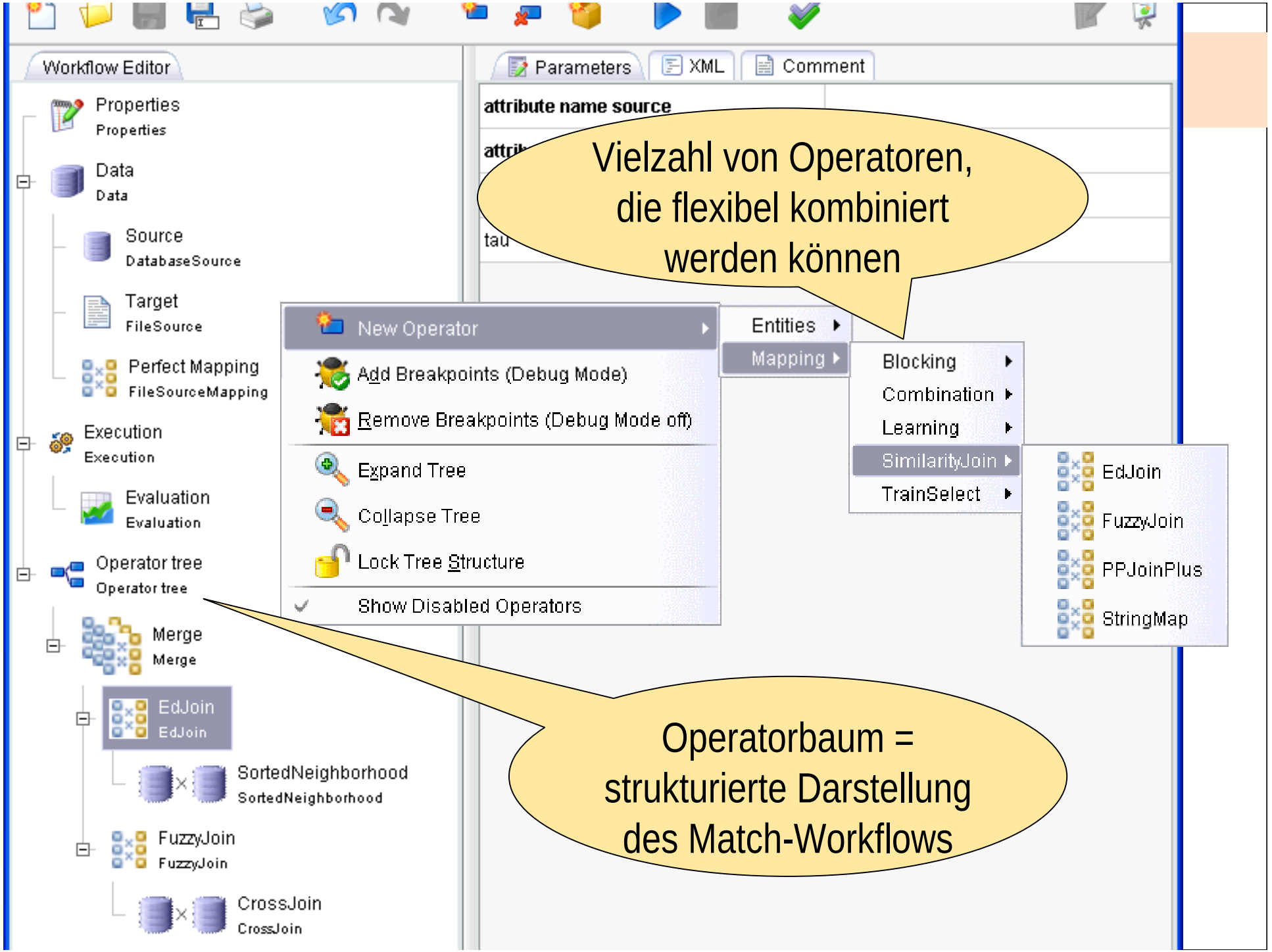
Objekt-Matching

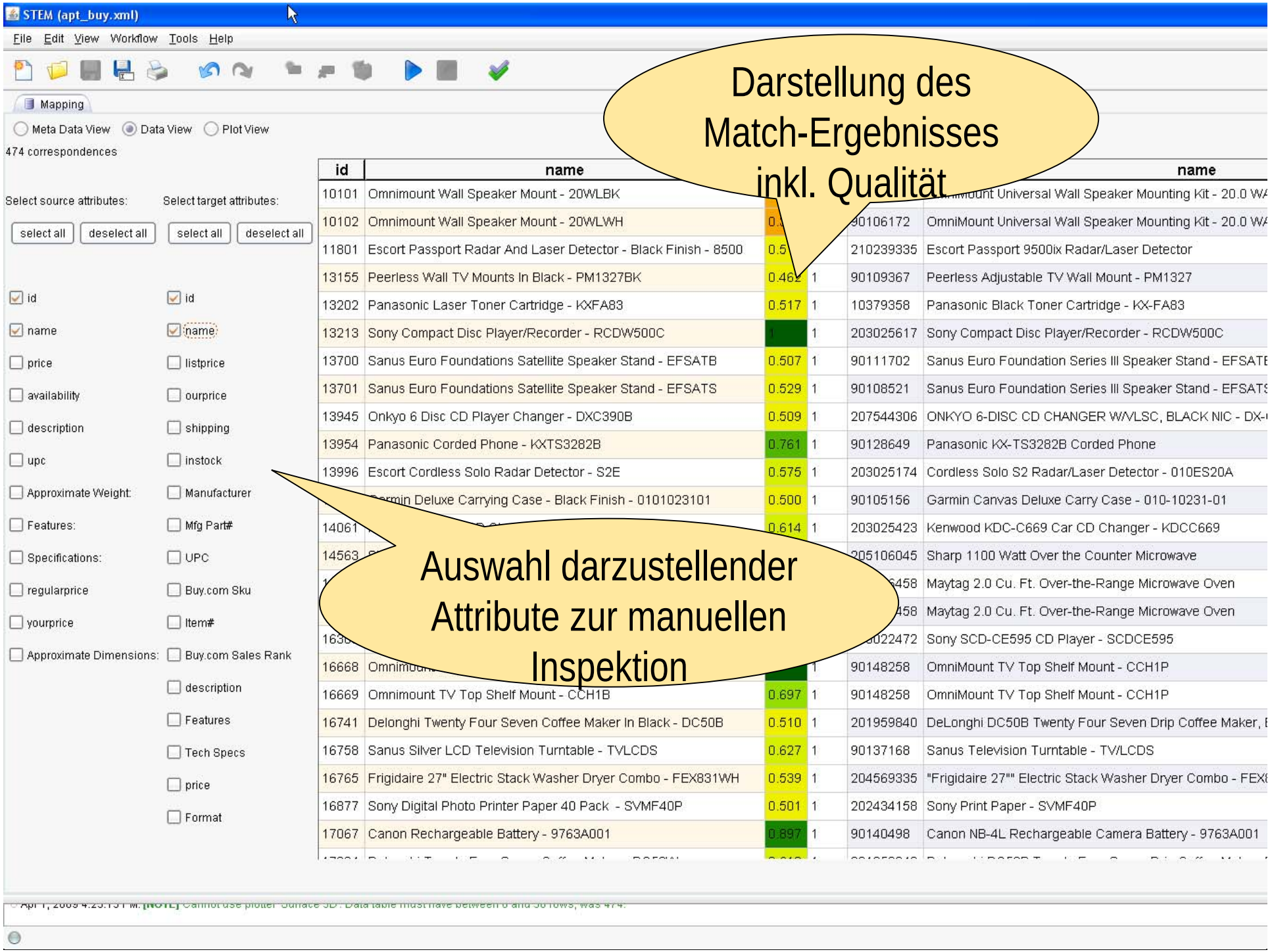
- Identifikation semantisch äquivalenter Objekte
  - innerhalb einer Datenquelle oder zwischen verschiedenen Quellen
  - um Objekte zu integrieren/mischen, zu vergleichen, Dubletten zu eliminieren, etc
- Anwendungsbereiche
  - Kunden-/Adressdaten
  - Produkte
  - Geografische Orte
  - ...
- Forschungsarbeiten seit 2006
  - Prototypen: MOMA und die Weiterentwicklung STEM

Prototypen MOMA und STEM: Features

- Erweiterbare Bibliothek von Match-Verfahren (Matcher)
  → Anpassbar an verschiedene Szenarien / Domänen
- Mapping-Kombination
  → Effiziente Berechnung
  → Qualitätsabsicherung / -steigerung
- Konstruktion von Match-Workflows
  → an Problem angepasste Lösungsstrategie
- Speichern von Mappings in Repository
  → Wiederverwendung von Match-Ergebnissen
- Automatische Einstellung relevanter Parameter
  → Verringerung des Konfigurationsaufwandes

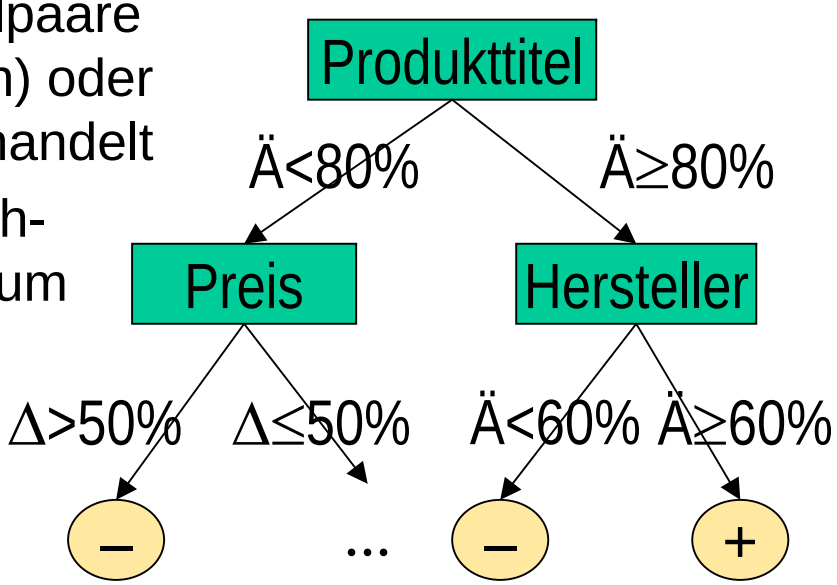

STEM
File Edit View Workflow Tools Help
Workflow Editor
Parameters
XML
Comment
Properties
Properties
Data
Data
Source
Target
Execution
Operator tree
Start Data Loading Wizard...
MySQL
database_url
username
Edit Text...
Choose source
Add Breakpoints (Debug Mode)
Remove Breakpoints (Debug Mode
Expand Tree
Collapse Tre
Lock Tree Structure
Show Disabled Operators
DatabaseSource
FileSource
QuerySource
Definition der Match-Aufgabe in Workflow-Editor
Unterstützung unter-schiedlicher Eingabequellen
5:00:41 PM


Workflow Editor
Properties
Properties
Data
Data
Source
DatabaseSource
Target
FileSource
Perfect Mapping
FileSourceMapping
Execution
Execution
Evaluation
Evaluation
Operator tree
Operator tree
Merge
Merge
EdJoin
EdJoin
SortedNeighborhood
SortedNeighborhood
FuzzyJoin
FuzzyJoin
CrossJoin
CrossJoin
Parameters
XML
Comment
attribute name source
New Operator
Add Breakpoints (Debug Mode)
Remove Breakpoints (Debug Mode off)
Expand Tree
Collapse Tree
Lock Tree Structure
Show Disabled Operators
Entities
Mapping
Blocking
Combination
Learning
SimilarityJoin
TrainSelect
EdJoin
FuzzyJoin
PPJoinPlus
StringMap
Vielzahl von Operatoren, die flexibel kombiniert werden können
Operatorbaum = strukturierte Darstellung des Match-Workflows


STEM (apt_buy.xml)
File Edit View Workflow Tools Help
Mapping
Meta Data View
Data View
Plot View
474 correspondences
Darstellung des Match-Ergebnisses inkl. Qualität
Select source attributes:
Select target attributes:
select all
deselect all
select all
deselect all
id
name
price
availability
description
upc
Approximate Weight:
Features:
Specifications:
regularprice
yourprice
Approximate Dimensions:
id
name
listprice
ourprice
shipping
instock
Manufacturer
Mfg Part#
UPC
Buy.com Sku
Item#
Buy.com Sales Rank
description
Features
Tech Specs
price
Format
Auswahl darzustellender Attribute zur manuellen Inspektion
id
name
name
10101 Omnimount Wall Speaker Mount - 20WLBK
10102 Omnimount Wall Speaker Mount - 20WLWH 90106172 OmniMount Universal Wall Speaker Mounting Kit - 20.0 WA
11801 Escort Passport Radar And Laser Detector - Black Finish - 8500 0.5 210239335 Escort Passport 9500ix Radar/Laser Detector
13155 Peerless Wall TV Mounts In Black - PM1327BK 0.462 1 90109367 Peerless Adjustable TV Wall Mount - PM1327
13202 Panasonic Laser Toner Cartridge - KXFA83 0.517 1 10379358 Panasonic Black Toner Cartridge - KX-FA83
13213 Sony Compact Disc Player/Recorder - RCDW500C 1 1 203025617 Sony Compact Disc Player/Recorder - RCDW500C
13700 Sanus Euro Foundations Satellite Speaker Stand - EFSATB 0.507 1 90111702 Sanus Euro Foundation Series III Speaker Stand - EFSATB
13701 Sanus Euro Foundations Satellite Speaker Stand - EFSATS 0.529 1 90108521 Sanus Euro Foundation Series III Speaker Stand - EFSATS
13945 Onkyo 6 Disc CD Player Changer - DXC390B 0.509 1 207544306 ONKYO 6-DISC CD CHANGER W/VLSC, BLACK NIC - DX-
13954 Panasonic Corded Phone - KXTS3282B 0.761 1 90128649 Panasonic KX-TS3282B Corded Phone
13996 Escort Cordless Solo Radar Detector - S2E 0.575 1 203025174 Cordless Solo S2 Radar/Laser Detector - 010ES20A
Garmin Deluxe Carrying Case - Black Finish - 0101023101 0.500 1 90105156 Garmin Canvas Deluxe Carry Case - 010-10231-01
14061 0.614 1 203025423 Kenwood KDC-C669 Car CD Changer - KDCC669
14563 205106045 Sharp 1100 Watt Over the Counter Microwave
6458 Maytag 2.0 Cu. Ft. Over-the-Range Microwave Oven
458 Maytag 2.0 Cu. Ft. Over-the-Range Microwave Oven
022472 Sony SCD-CE595 CD Player - SCDCE595
16668 Omnimoun 1 90148258 OmniMount TV Top Shelf Mount - CCH1P
16669 Omnimount TV Top Shelf Mount - CCH1B 0.697 1 90148258 OmniMount TV Top Shelf Mount - CCH1P
16741 Delonghi Twenty Four Seven Coffee Maker In Black - DC50B 0.510 1 201959840 DeLonghi DC50B Twenty Four Seven Drip Coffee Maker,
16758 Sanus Silver LCD Television Turntable - TVLCDS 0.627 1 90137168 Sanus Television Turntable - TV/LCDS
16765 Frigidaire 27" Electric Stack Washer Dryer Combo - FEX831WH 0.539 1 204569335 "Frigidaire 27"" Electric Stack Washer Dryer Combo - FEX
16877 Sony Digital Photo Printer Paper 40 Pack - SVMF40P 0.501 1 202434158 Sony Print Paper - SVMF40P
17067 Canon Rechargeable Battery - 9763A001 0.897 1 90140498 Canon NB-4L Rechargeable Camera Battery - 9763A001

Automatische Match-Konfiguration (Self-Tuning)

- Richtige Match-Strategie: Fragestellungen
  - Welche (relevanten) Attribute sollen verglichen?
  - Wie sollen die Ähnlichkeitswerte kombiniert / verrechnet werden?
- Automatische Konfiguration mittels Lernverfahren
  - Nutzer gibt für (wenige!) Beispielpaare an, ob es sich um gleiche (match) oder ungleiche (non-match) Objekte handelt
  - System errechnet optimale Match-Strategie, z.B. Entscheidungsbaum



Feedback / Diskussion

- Wie beurteilen Sie die vorgestellten Verfahren (Stärken, Schwachstellen, Erweiterungsbedarf)?
- Sehen Sie Anwendungsmöglichkeiten für die vorgestellten Verfahren?
  - zur Lösung konkreter Integrationsaufgaben
  - zur Kombination mit eigenen Entwicklungen/Produktangeboten
- Wie schätzen Sie die Marktrelevanz der Lösungsansätze ein?
- Welche Kooperationsformen mit dem Innovationslabor können Sie sich vorstellen, z.B.
  - gemeinsam betreute studentische Arbeiten
  - gemeinsame Projekte z.B. zur Pilotanwendung der Werkzeuge
  - Beratung zu Datenintegrationsthemen