Universität Leipzig
Institut für Informatik

Auffinden von Dubletten in ECommerce Datenbeständen

Hanna Köpcke
AG 3: Objekt Matching

06.05.2010

Agenda

- Problemstellung
- FEVER-System
  - Manuell definierte Match-Strategien
  - Trainingsbasierte Match-Strategien
  - Evaluierung
- Anwendungsszenarien
- Zusammenfassung

Erkennung von Dubletten (Objekt-Matching)

- Identifikation semantisch äquivalenter Objekte
  - z.B. zur Eliminierung, Fusion oder zum Datenvergleich
  - kritischer Schritt für hohe Datenqualität
- derzeit v.a. für strukturierte (relationale) Daten

Quelle1: Kontakt

KID	Name	Strasse	Stadt	Frau
11	Kristen Schmid	Hanse Pl 2	Berlin	1
24	Christian Schmied	Hanse Str 2	Berlin	0

Quelle2: Kunde

Kdnr	Nachname	Vorname	Geschl	Adresse	Telefon
11	Schmid	Chris	M	Hansestr. 2, 18182 Bentwich	
493	Schmid	Kris L.	W	Hansa-Platz 2, 10557 Berlin	030-9627621

Dubletten in Ecommerce Webdaten

Microsoft Xbox 360 Arcade
Inklusive: UNO, Pac-Man Championship Edition, Luxor 2, Feeding Frenzy, Boom Boom Rocket
Das **Xbox 360 Arcade** System bringt alles mit, was ihr braucht, um mit dem Spielen loszulegen. Der Zeitpunkt zum Einsteigen war noch nie so günstig: ihr bekommt den Wireless ...
Zur Einkaufsliste hinzufügen

€115 neu, €109 gebraucht
von 91 Händlern
Preise vergleichen

Microsoft Xbox 360 Arcade
... Composite-AV-Kabel mit SCART Adapter, **Xbox** Live **Arcade** Compilation, inkl. Banjo Kazooie Weitere Infos **Xbox 360**, das einzige Videospielsystem der **...**.
Zur Einkaufsliste hinzufügen

€149,00 neu
€158,90 mit Versand
groundPC Computersysteme

Microsoft XBox 360 Arcade
Microsoft **XBox 360 Arcade**: Das **Xbox 360 Arcade** System bringt alles mit, was ihr braucht, um mit dem Spielen loszulegen. Der Zeitpunkt zum Einsteigen war **...**.
Zur Einkaufsliste hinzufügen

€199,90 neu
€203,85 mit Versand
Passiontec.de (günstigste Versandart)
★★★★★ 149 Händlerbewertungen

Microsoft XBox 360 Arcade und HTC Hero White und LG KP100 Black und Ap
Top-Zugabe: Microsoft **XBox 360 Arcade** uvm. nach Wahl geschenkt Talkline Extra Duo Tarif Je 50 Freiminuten monatl. inklusive! Nur je 14,95 Euro monatlicher **...**.
Zur Einkaufsliste hinzufügen

€1,00 neu
Kostenloser Versand
eteleon Handy Shop
★★★★★ 1 Händlerbewertung

xbox 360 arcade system
Ausbaufähig: **Xbox 360 Arcade** ist mit Zubehör wie **Xbox 360**-Festplatten AV-Kabeln Memory Units Wireless Controllern und vielen weiteren interessanten **...**.
Zur Einkaufsliste hinzufügen

€199,00 neu
Kostenloser Versand
Versandkosten inklusive - Gimahhot

, XBox360 Real Arcade Pro 3 Fighting Stick
XBox360 Real **Arcade** Pro 3 Fighting StickDas große Gehäuse und das hohe Gewicht von ca. 2,6 kg geben dem neuen **Arcade** Stick eine gesteigerte Stabilität, **...**.
Zur Einkaufsliste hinzufügen

€129,95 neu
€135,90 mit Versand
Conrad Electronic
★★★★½ 173 Händlerbewertungen

Herausforderungen

- Sehr ähnliche Attributwerte (Title, Beschreibung, Preis, ..) für ähnliche aber unterschiedliche Produkte
- Heterogene Repräsentationen für das gleiche Produkt
- Geringe Datenqualität
  - Fehlende Angaben
  - Fehlerhafte Angaben

Inakustik Star Lautsprecherkabel
Star Lautsprecherkabel 2 x 1,5 mm², transparent, Länge 10m

≠

Inakustik Star Lautsprecherkabel
Star Lautsprecherkabel 2 x 2,5 mm², transparent, Länge 10m

Nikon Blitzgerät SB-900

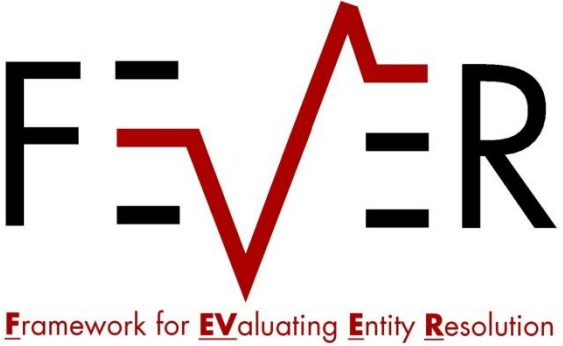
=

NIKON Speedlight SB-900 Leitzahl 34 48 Blitzausleuchtung Brennweite 17-200mm

Objekt-Matching-Ansätze

- Zahlreiche Forschungsansätze und -prototypen sowie kommerzielle Lösungen
- Zumeist Nutzung von Ähnlichkeiten von Attributwerten
  - z.B. gemäß String-Ähnlichkeitsmaßen
- Probleme
  - Effektive Kombination mehrerer Match-Verfahren
  - Hoher Tuning-Aufwand für Konfigurierung (z.B. Auswahl relevanter Attribute, Ähnlichkeitsschwellwerte, Gewichtung einzelner Verfahren)
  - Laufzeit für große Datenmengen

FEVER Framework



- FEVER = **F**ramework for **EV**aluating **E**ntity **R**esolution
  - System zur Definition, Konfigurierung und Evaluierung von Objekt-Matching (entity resolution)-Strategien
- Wesentliche Merkmale:
  - Flexible Kombination mehrerer Match-Verfahren im Rahmen von Objekt-Matching-Workflows
  - Semi-automatische Parameter-Konfigurierung, z.B. für Ähnlichkeitsschwellwerte
  - Unterstützung trainingsbasierter Match-Verfahren zur Reduzierung des manuellen Tuningaufwands
  - Vergleichende Analyse alternativer Verfahren

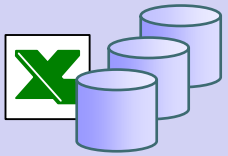
FEVER Architektur

GUI

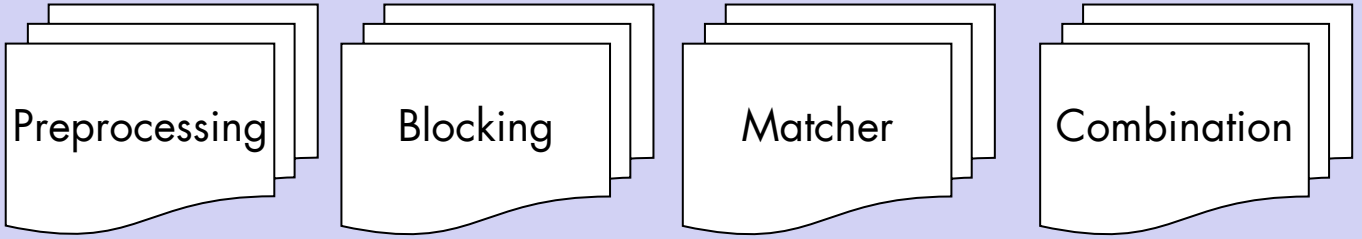
Workflow Definition

Optimization

Workflow Execution Engine



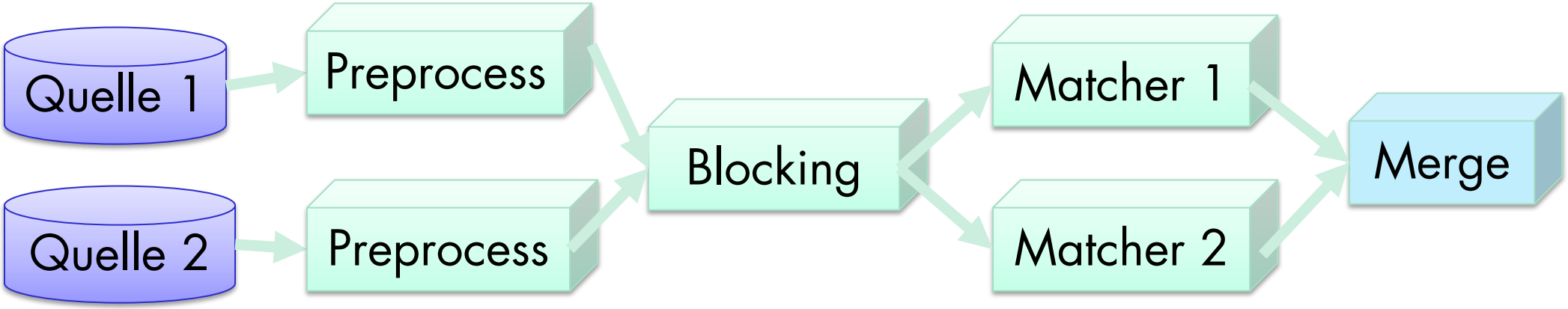
Data Services



Operator Library

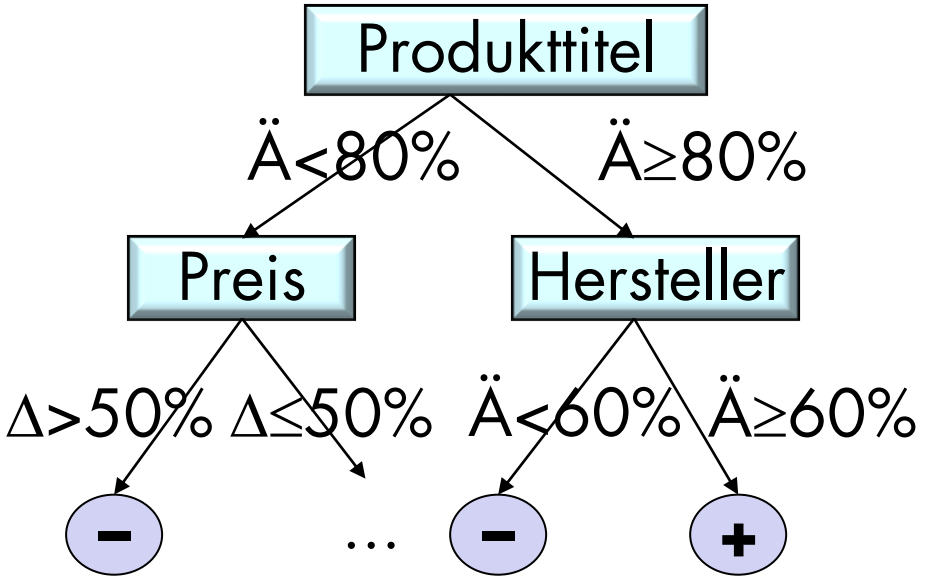
Match Workflow

- Vorverarbeitung
- Blocking zur Reduzierung des Suchraumes
  - z.B. durch Clustering, Sorted Neighborhood
- Attribut-Matcher sowie Kontext-Matcher
  - zahlreiche Ähnlichkeitsfunktionen und externe Implementierungen



Trainingsbasierte Strategien

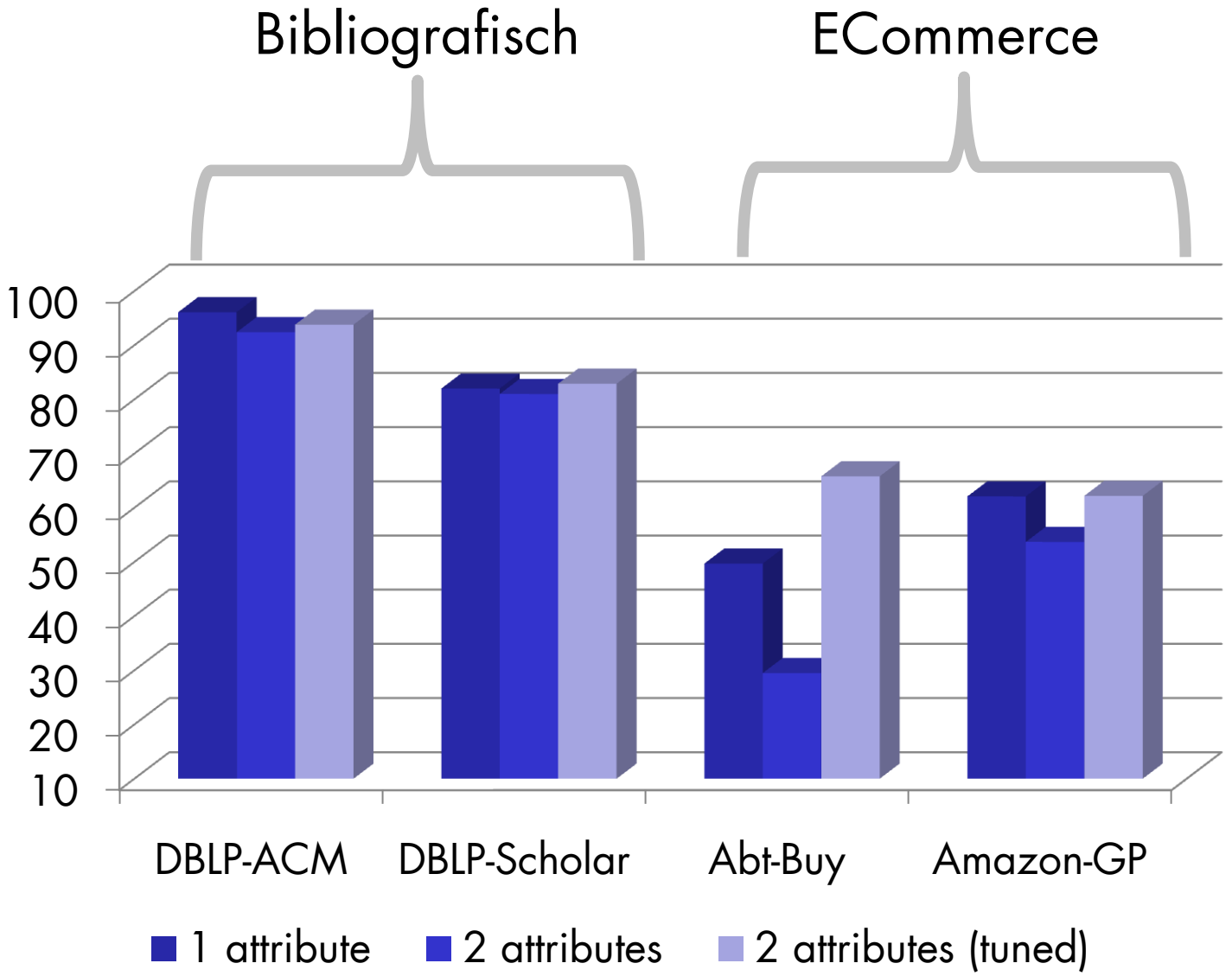
- Nutzung von Trainingsdaten um effektive Kombination von Matchern und deren Konfigurierung zu bestimmen (supervised learning)
- In FEVER unterstützte Lernverfahren:
  - Entscheidungsbaum, Logistische Regression, SVM
  - Mehrheits-Lerner



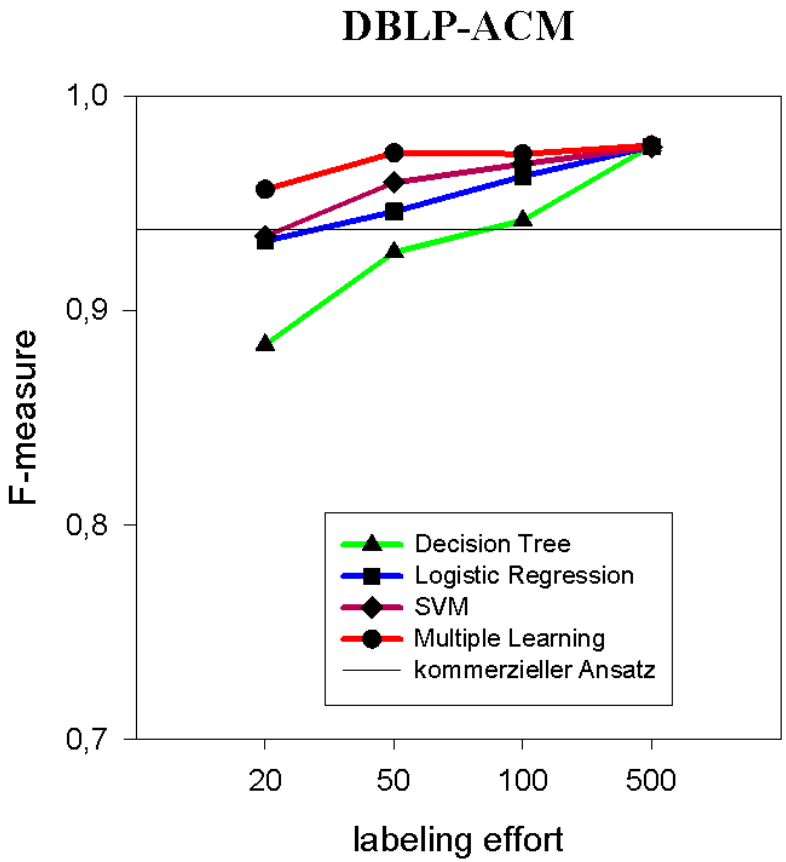
Evaluation

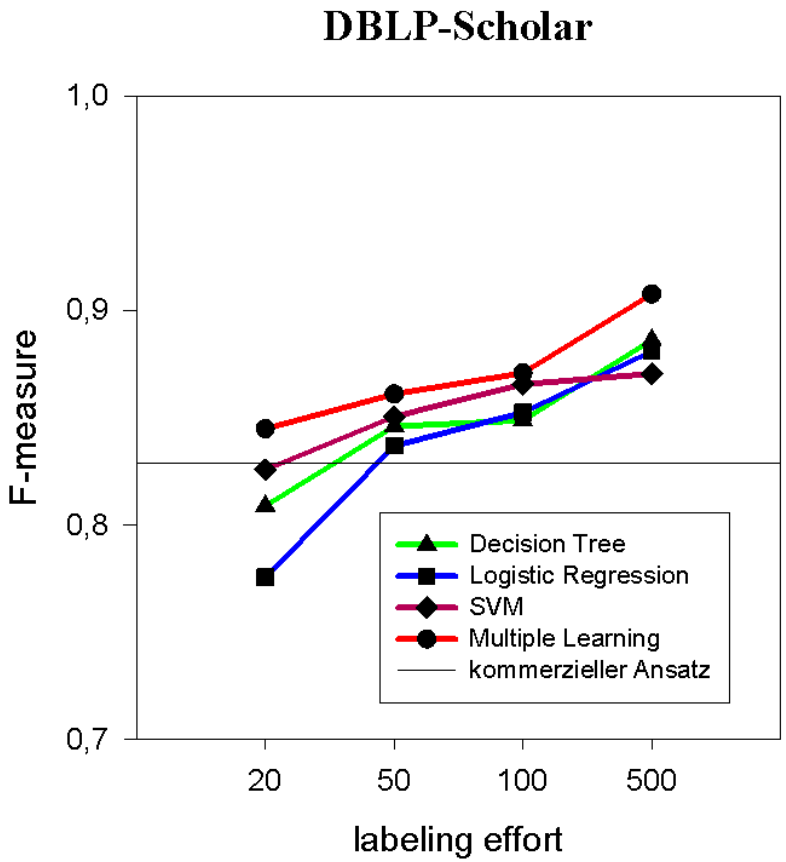
- 4 Matchaufgaben mit 7 Datenquellen
  - bibliographisch: DBLP-ACM
    DBLP-Google Scholar (GS)
  - E-Commerce: Abt-Buy
    Amazon - GP
- bis zu 64,000 Objekte pro Quelle
- Perfektes Ergebnis bekannt
  - Manuell bestimmt bzw. über UPCs für Produktdaten
- Vergleich zwischen
  - kommerziellem Match-Ansatz mit Parameteroptimierung und trainingsbasierten Ansätzen

Tuning des kommerziellen Match-Ansatzes

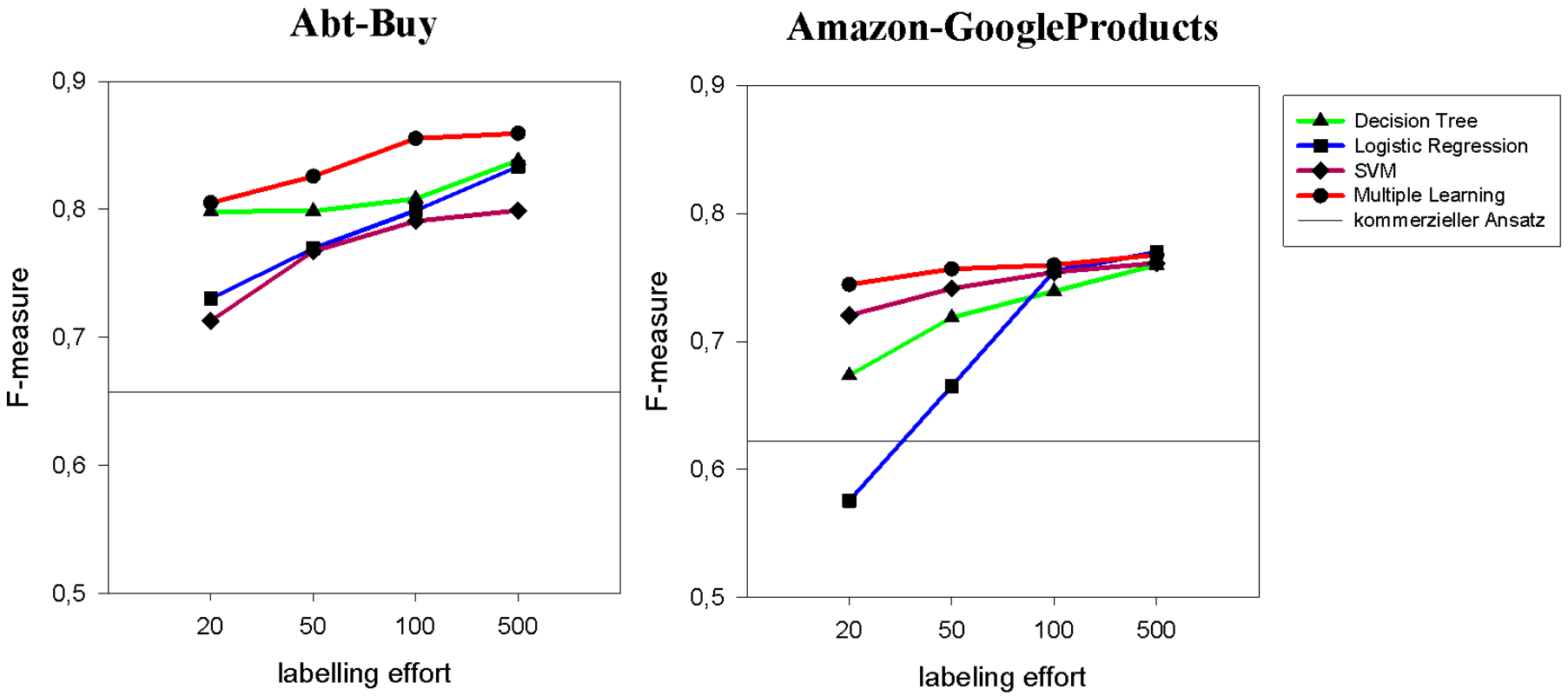


Ergebnisse Matching von Publikationen





Ergebnisse Matching von Produkten



Anwendungsszenarien

- Integration und Aufbereitung unternehmensinterner und externer Daten (Webdaten) für weitergehende Analysen, z.B. für
  - Customer Relationship Management
  - Kundenbewertung analysieren
  - Erstellung von Konkurrenzanalysen (Produkt, Preis, Anbieter, Zielgruppen)
  - …

Zusammenfassung

- Flexible Kombination mehrerer Match-Verfahren
- Semi-automatische Parameter-Konfigurierung, auch für externe Matchansätze
- Unterstützung trainingsbasierter Match-Verfahren zur Reduzierung des manuellen Tuningaufwands
- Gute Effektivität für bibliografische Probleme
  - F-Measure > 91%
- E-Commerce Daten deutlich schwieriger
  - F-Measure 77-86%
- Vergleich mit kommerzieller Lösung
  - Bis zu 15% höhere Performanz (F-Measure)

Vielen Dank für Ihre Aufmerksamkeit!