Universität Leipzig
Institut für Informatik

Dublettenerkennung und Bereinigung von Webdaten

Hanna Köpcke
WDI-Lab

28.10.2010

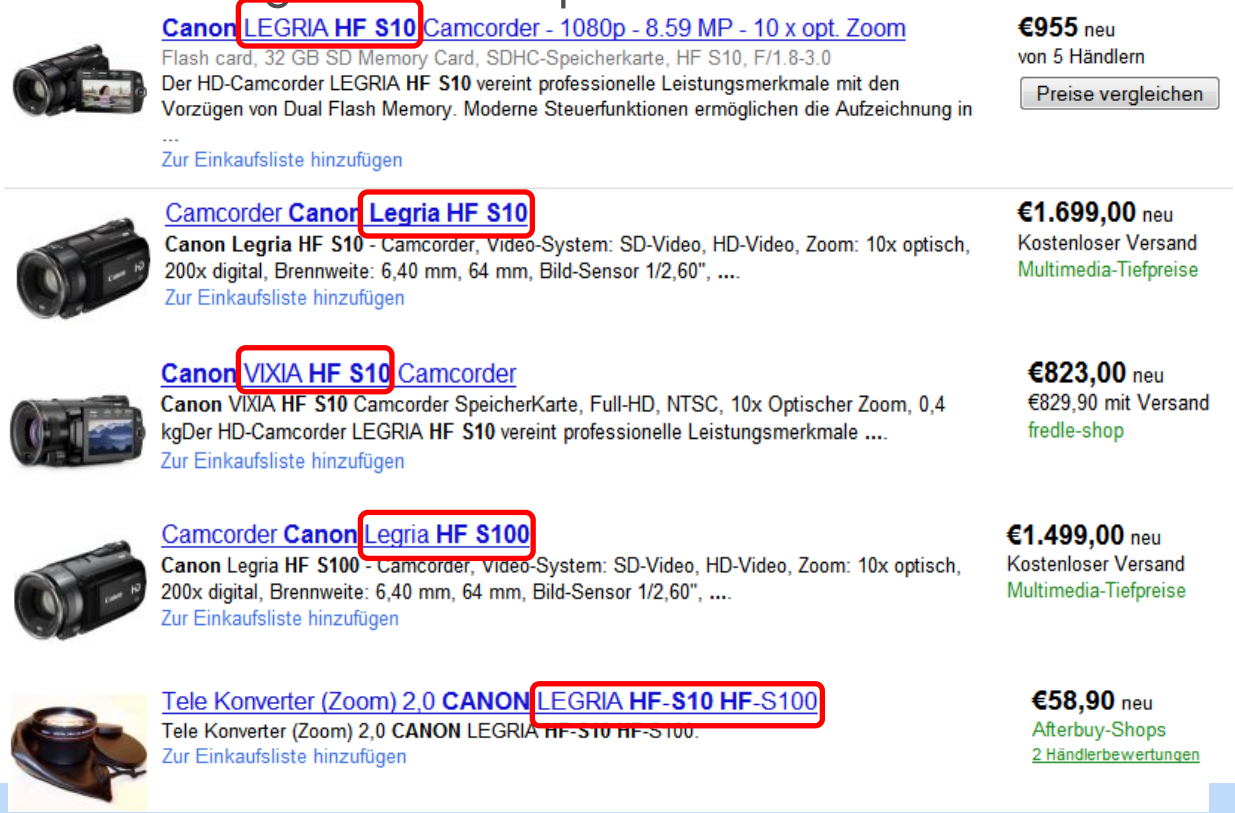
Herausforderungen für die Webdatenintegration

- Schlechte Datenqualität
  - Heterogene Repräsentationen
  - Fehlende Angaben
  - Fehlerhafte Angaben
- Große Datenmengen
- Verarbeitung in Echtzeit

Titel	Hersteller	Eigenschaften
Taschenradio ICF-304	Taschenradio ICF-304	silber
Sony ICF-304 tragbares Radio (38459)	Sony	
HP Photosmart C6180	HP	weiß
HP Photosmart C6180	Hewlett Packard	schwarz
Kyocera FS 1300D	Kyocera	
C7280 inkl. Ersatzpatronen		
Kyocera FS 1300D	Kyozera	

Erkennung von Dubletten

- Identifikation semantisch äquivalenter Objekte
  - z.B. zur Eliminierung, Integration oder zur Analyse
  - Erhöhung der Datenqualität



Canon LEGRIA HF S10 Camcorder - 1080p - 8.59 MP - 10 x opt. Zoom
Flash card, 32 GB SD Memory Card, SDHC-Speicherkarte, HF S10, F/1.8-3.0
Der HD-Camcorder LEGRIA **HF S10** vereint professionelle Leistungsmerkmale mit den Vorzügen von Dual Flash Memory. Moderne Steuerfunktionen ermöglichen die Aufzeichnung in ...
Zur Einkaufsliste hinzufügen
€955 neu
von 5 Händlern
Preise vergleichen

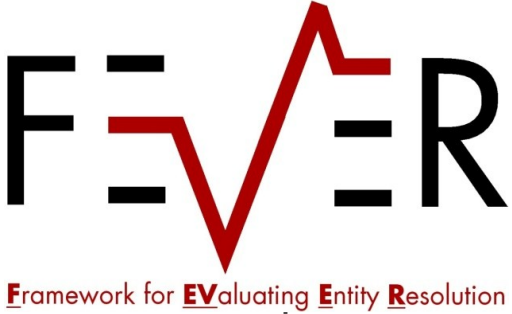
Camcorder **Canon Legria HF S10**
Canon Legria HF S10 - Camcorder, Video-System: SD-Video, HD-Video, Zoom: 10x optisch, 200x digital, Brennweite: 6,40 mm, 64 mm, Bild-Sensor 1/2,60", ...
Zur Einkaufsliste hinzufügen
€1.699,00 neu
Kostenloser Versand
Multimedia-Tiefpreise

Canon VIXIA HF S10 Camcorder
Canon VIXIA **HF S10** Camcorder SpeicherKarte, Full-HD, NTSC, 10x Optischer Zoom, 0,4 kgDer HD-Camcorder LEGRIA **HF S10** vereint professionelle Leistungsmerkmale ....
Zur Einkaufsliste hinzufügen
€823,00 neu
€829,90 mit Versand
fredle-shop

Camcorder **Canon** Legria **HF S100**
Canon Legria **HF S100** - Camcorder, Video-System: SD-Video, HD-Video, Zoom: 10x optisch, 200x digital, Brennweite: 6,40 mm, 64 mm, Bild-Sensor 1/2,60", ...
Zur Einkaufsliste hinzufügen
€1.499,00 neu
Kostenloser Versand
Multimedia-Tiefpreise

Tele Konverter (Zoom) 2,0 **CANON** LEGRIA **HF-S10 HF**-S100
Tele Konverter (Zoom) 2,0 **CANON** LEGRIA **HF-S10 HF**-S100.
Zur Einkaufsliste hinzufügen
€58,90 neu
Afterbuy-Shops
2 Händlerbewertungen

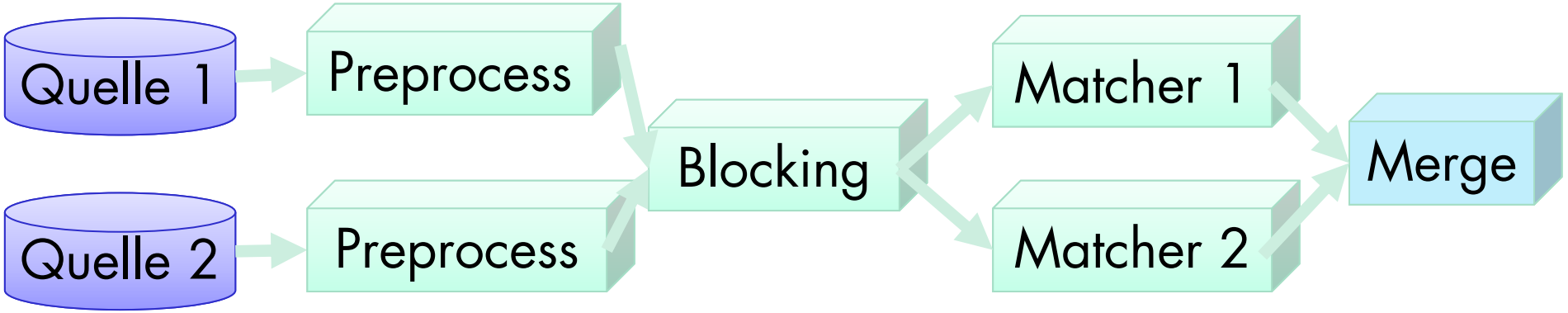
(Semi-) automatisches Matching



- Forschungsprototyp FEVER = Framework for EValuating Entity Resolution
  - System zur Definition, Konfigurierung und Evaluierung von Objekt-Matching (entity resolution)-Strategien
- Bibliotheken für
  - Effizienten Datenzugriff (Unterstützung unterschiedlicher Typen und Formate)
  - Vorverarbeitung (Bereinigung und Anreicherung von Webdaten)
  - Match-Verfahren
- 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

Beispielworkflow für die Identifizierung von Dubletten

- Vorverarbeitung
  - Datenbereinigung: Vereinheitlichung heterogener Angaben, Ergänzung fehlender Angaben, Informationsanreicherung
- Blocking zur Reduzierung des Suchraumes
  - z.B. durch Clustering, Sorted Neighborhood
- Attribut-Matcher sowie Kontext-Matcher
  - zahlreiche Ähnlichkeitsfunktionen und externe Implementierungen

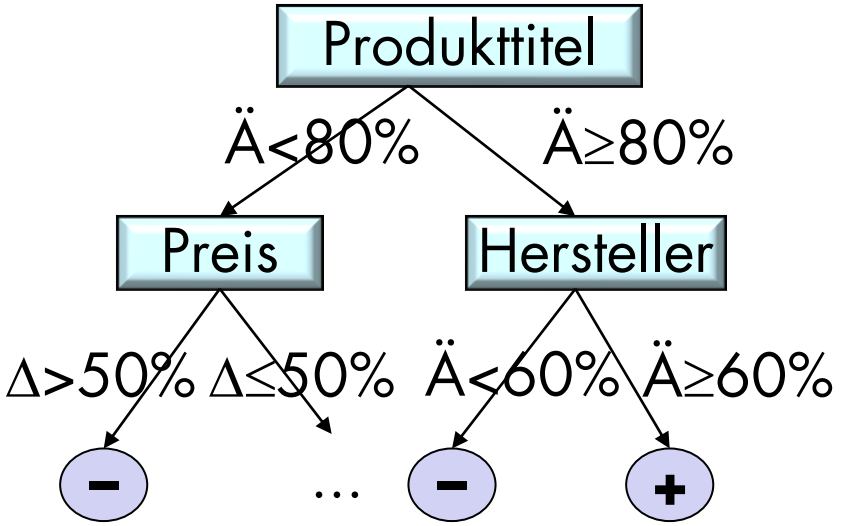


Daten-Bereinigung

- Iteratives Cleaning von Herstellerangaben
- Extraktion von Produkt-Codes
- Anreicherung von EANs
- Signifikante Qualitätsverrbesswerungen

Trainingsbasierte Strategien zur Erkennung von Dubletten

- Nutzung von Trainingsdaten um effektive Kombination von Matchern und deren Konfigurierung zu bestimmen (supervised learning)



Vorteile unserer Technologien

- (Semi-)automatisches Matching
- Strategien für große Datenmengen
- Matchen von Produktangeboten
  - zu Produkten in einem Produktkatalog
  - mit anderen Angeboten
- Aufbau /Erweiterung eines Produktkatalogs
- Erfolgreich eingesetzt auch für andere Daten
- Vergleich mit kommerzieller Lösung
  - 15% höhere Match-Qualität (F-Measure) bei E-Commerce-Daten

Vielen Dank für Ihre Aufmerksamkeit!