

Universität Leipzig
Institut für Informatik



Methoden zur Kategorisierung und Duplikaterkennung von Produktdaten

Hanna Köpcke

Web Data Integration Workshop 2011



Was sind die Probleme?

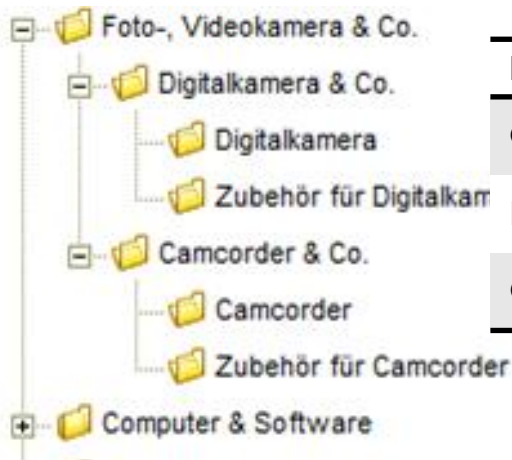
| Name | Händlerkategorie | Hersteller |
|--|------------------------------|--------------------|
| IXUS 115 HS - Digitalkamera - unterstützter Speicher: SD, SDXC, SDHC | Foto & Video | Canon |
| IXUS 115 HS Rosa | Kompaktkameras Einsteiger | Canon |
| Canon IXUS 115 HS, 12 Megapixel, Pink | Digitalkamera > Canon | Canon |
| Hähnel HL-XF51 7.2V 680mAh für Sony NP-FF51 | Kamera-Akku | Kamera-Akku |
| HP Officejet 4500 Multifunktionsgerät mit Fax | Multifunktionsgerät | Hewlett Packard |
| Officejet 4500 | | HP |

Lösungen

Kategorisierung

Datenbereinigung

Duplikaterkennung



| Name | Hersteller | Code |
|---------------------|-----------------|-------------|
| Canon IXUS 115 HS | Canon | IXUS 115 HS |
| Hähnel HL-XF51 7.2V | Hähnel Akku | HL-XF51 |
| Officejet 4500 | Hewlett Packard | |

IXUS 115 HS Rosa

IXUS 115 HS



| Detailinfos | |
|----------------------------|--|
| Anzeige | |
| Monitor: | |
| Displaygröße: | |
| Camera display resolution: | |
| Audio | |
| Eingebautes Mikrophon: | |
| Belichtung | |
| Auto exposure: | |
| ISO Empfindlichkeit: | |
| Kamera Verschlusszeit: | |
| Belichtungskorrektur: | |

Kategorie

| | |
|---------------------|-----------------------------------|
| IXUS 115 HS Rosa | Digitalkamera |
| IXUS 105 | Digitalkamera |
| Hähnel HL-XF51 7.2V | Zubehör für Digitalkameras |

IXUS 115 HS Rosa

IXUS 115 HS

IXUS 105

IXUS 105

Kategorisierung

- Herausforderungen:
 - Große Anzahl an Kategorien und Angeboten
 - Große Unterschiede zwischen Katalogen
 - Struktur, Bezeichnungen
 - „Falsche“ Zuordnungen
 - Besseres Placement (Zubehör in Hauptkategorie)
 - Mehrere mögliche Kategorisierungen

| | |
|----------------------|---|
| Identifizier: | 14978321 |
| Name: | HC-BM10 Drahtloses Babyfon weiß/blau |
| Beschreibung: | Alarm beim Verlassen des Empfangsbereichs LED |
| Hersteller: | König |
| Preis: | 65.95 |
| Bild: |  |

Babyartikel?
Funktechnik?

(Semi-)automatische Kategorisierung

① Training

Tokenisierung

Häufigkeitsverteilung

Lernen

② Anwendung

Tokenisierung

Kategorisierung

Modell

Anpassung des Modells

Inkrementelle Verbesserung

Falsch?

Top N Vorschläge

HC-BM10 Drahtloses Babyfon weiß/blau

Funktechnik & Zubehör (0,17)

| category | confidence |
|--------------------------------|------------|
| sonstige Baby- & Kinderartikel | 0.83 |
| Funktechnik & Zubehör | 0.17 |
| Baby- & Kinderpflege | 0.00 |
| Baby-Ernährung | 0.00 |
| Kinderwagen & Zubehör | 0.00 |

Datenbereinigung

- Cleaning kritischer Attribute, z.B. von Hersteller- oder Markenangaben

- Automatisierte Erstellung von Lookup-Tabellen mit Synonym-Verwaltung

| | |
|----------------|----------------|
| Allied Telesis | Allied Telesis |
| Allied | Allied Telesis |
| Allied Telesyn | Allied Telesis |

- Feature-Extraktion und -Verbesserung

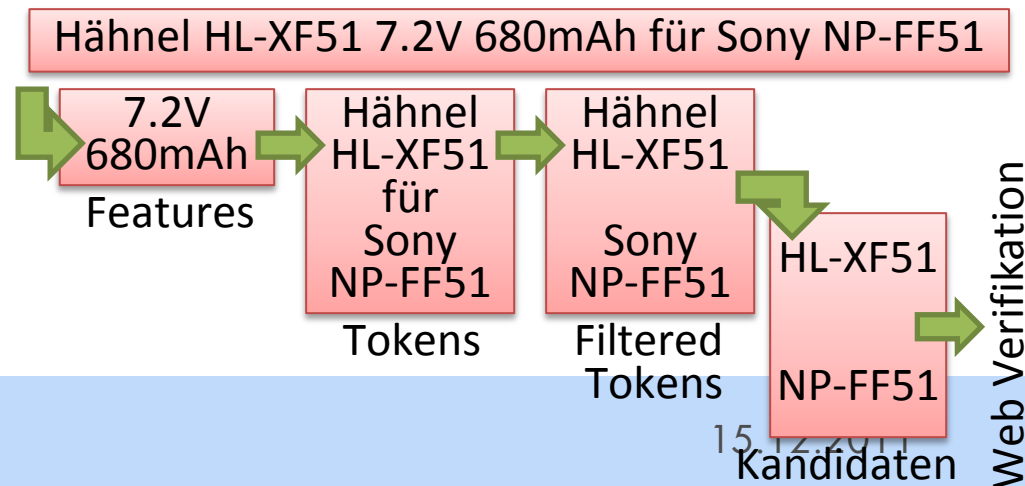
- Produkt-Ids, z.B. EAN, GTIN



HC-BM10 Drahtloses Babyfon weiß/blau

Alarm beim Verlassen des Empfangsbereichs LED EAN: 5412810142194

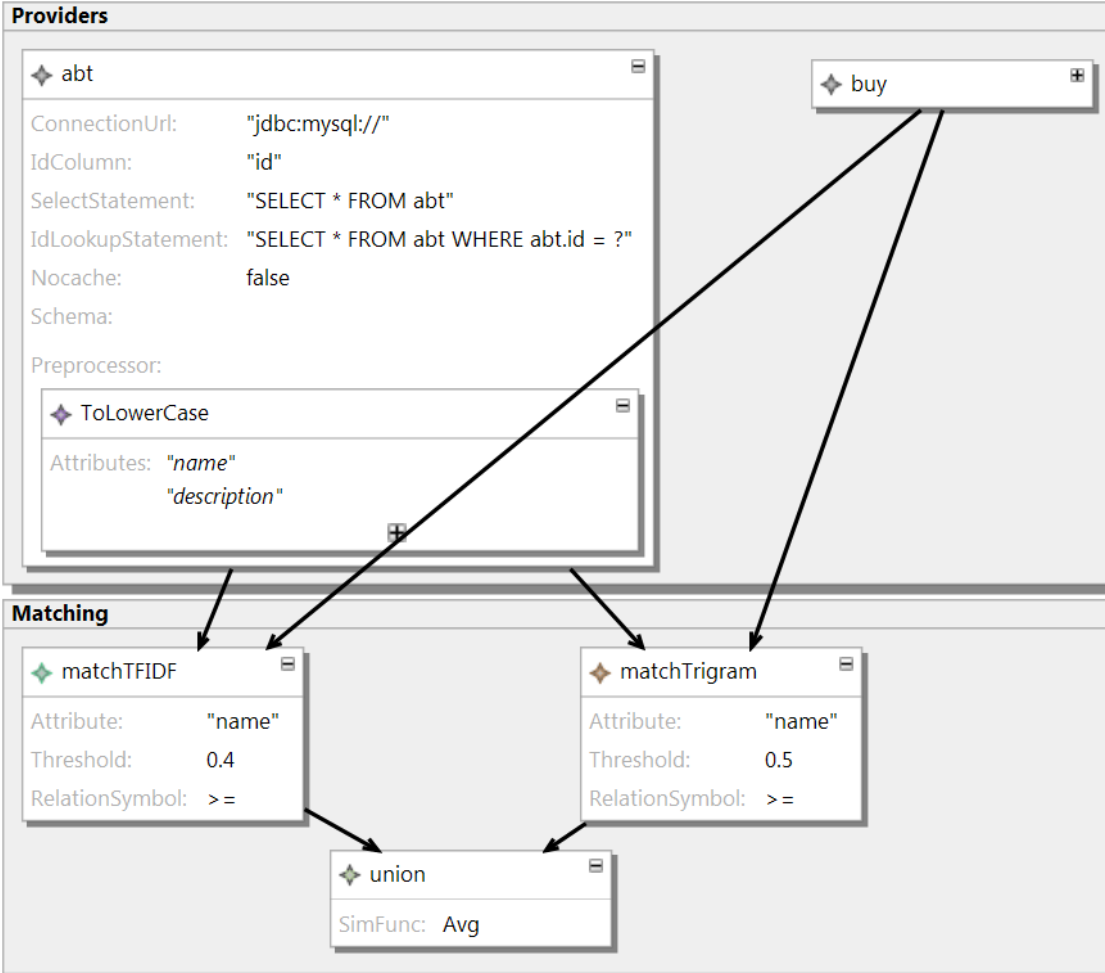
- Produkt-Codes



Matchen zur Erkennung von Duplikaten

- Matchen = Identifikation semantisch äquivalenter Objekte
- Matchen von Produktangeboten
 - zu Produkten in einem Produktkatalog
 - mit anderen Angeboten
- Flexible Kombination mehrerer Match-Verfahren im Rahmen von Objekt-Matching-Workflows

Match-Workflow



Trainingsbasierte Match-Verfahren

- Reduzierung des manuellen Tuning-Aufwandes
- Trainingsdaten
 - Unterstützung bei der manuellen Annotation



IXUS 115 HS Rosa

Digitale Kompaktkamera mit 12,1 Megapixel

IXUS 115 HS

Digitalkamera, Kompaktkamera, 12.1 Mpix,
4-fach optischer Zoom

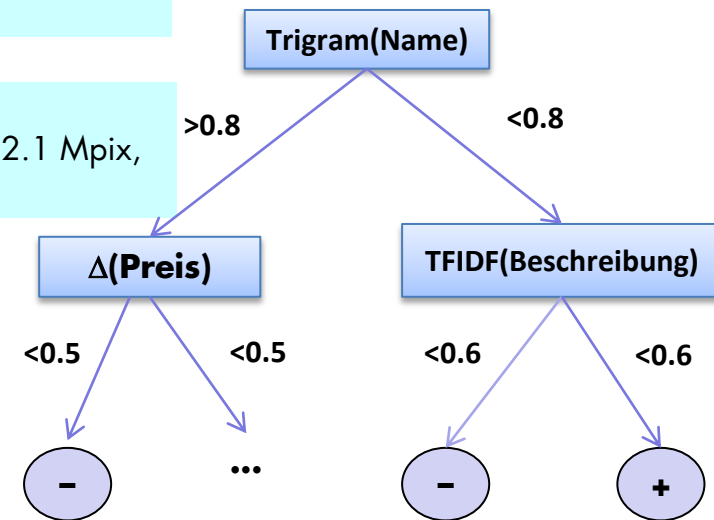
IXUS 115 HS

Digitalkamera, Kompaktkamera, 12.1 Mpix,
4-fach optischer Zoom

IXUS 105

Digitalkamera, Kompaktkamera, 12.1 Mpix,
4-fach optischer Zoom

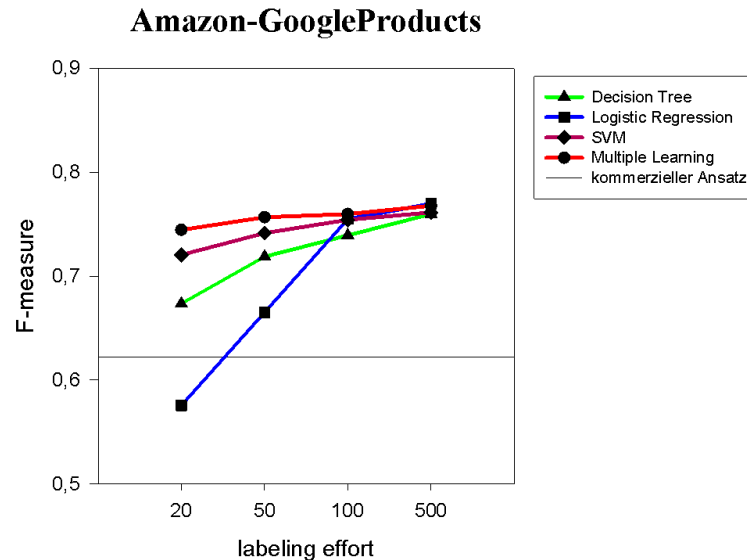
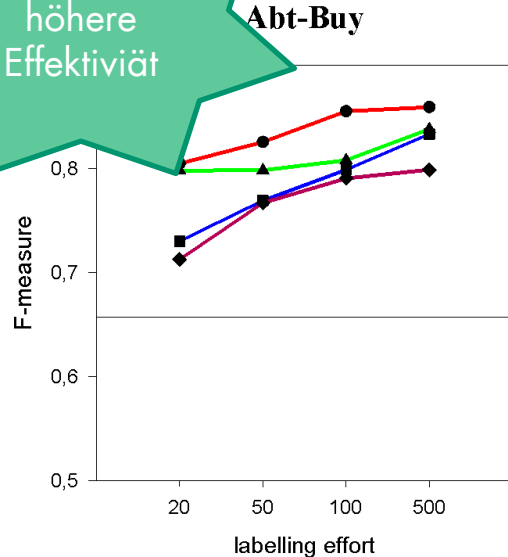
- Maschinelle Lernverfahren:
 - Entscheidungsbaum, SVM, ANN
 - Logistische Regression, Mehrheitslerner



Vorteile unserer Lösungen

- Erhöhung der Datenqualität
- Einsparung von Zeit und Kosten
- Optimiert für große Datenmengen
- Branchenübergreifend einsetzbar

Bis zu 15 %
höhere
Effektivität





Online Product Manager

Automatische Kategorisierung

Top N Vorschläge

File Categorise Match Filter Filter

tree repository category Tree

- Foto-, Videokamera & Co.
- Computer & Software
- Schnäppchen & Gebrauchtware
- Audio, Video & TV
- Buch, Hörbuch & Kalender
- Beauty, Wellness & Drogerieartikel
- Downloads zum Verkauf & Verleih
- Telekommunikation
- Game & Konsole
- Hobby & Spiel
- Haushalt & Wohnen
- Sport & Freizeit
- Mode & Accessoires
- weitere Produkte

product set repository product set

| number | Name | Beschreibung | Hersteller | Preis | category (confidence) |
|--------|--|---|------------|--------|---------------------------------------|
| 0 | HC-BM50 2,4 GHz Funk-Babyfon mit Digitalkamera weiß/blau | Keine Einmischung und höchst privaten Automatische infrarot night v König | | 115.23 | Funktechnik & Zubehör (0,76) |
| 0 | HC-BM10 Drahtloses Babyfon weiß/blau | Alarm beim Verlassen des Empfangsbereichs LED Lautstärkeanzeige König | | 65.95 | sonstige Baby- & Kinderartikel (0,83) |
| 0 | HC-BM10 Drahtloses Babyfon weiß/blau | KÖNIG DRAHTLOSES BABYPHONE Legen Sie Ihr Baby schlafen und König | | 65.95 | sonstige Baby- & Kinderartikel (0,83) |
| 0 | Walker 6800 PMR Funkgeräte mit beleuchtetem LC-Display und bis z Verpackungseinheit: 1 Stück. beleuchtetes LC-Display, VOX-Funktion Topcom | | | 54.98 | sonstige Baby- & Kinderartikel 0.83 |
| 0 | Cobra M... MP Walkie Talkies | Das neueste Gerät der Cobra Serie, Set mit 2 PMR-Geräten, 1 Ladegerät Unbekannt | | 53.49 | Funktechnik & Zubehör 0.17 |
| 2312 | Cobra MT600-2 V... Walkie Talkies | ... Akkus - 8 Kanäle - 3 Unbekannt | | 48.81 | Baby- & Kinderpflege 0.00 |
| 0 | MT200-2 VP Walkie Talkies | ... germodell. Komplettpaket COBRA | | 44.29 | Baby-Ernährung 0.00 |
| 0 | A010001 - Hochstuhl Slim ab 6 Mon... | ... gbarer und einstellbare Babymoov | | 109.99 | Kinderwagen & Zubehör 0.00 |
| 0 | 71291 - Babyphone - Tragetasche für... | ... D und SR Serie Tomy | | 3.0 | Baby-Ernährung (0,70) |
| 0 | 651600000 - Babyphone Bay Contro... | ... Classic Chicco | | 57.05 | Funktechnik & Zubehör (0,97) |
| 0 | 651600000 - Babyphone Bay Contro... | ... s, technisch-klares Det Chicco | | 57.99 | Funktechnik & Zubehör (0,79) |
| 0 | 55012 - Walkie Talkie | Hier spricht Spiderman, over. Mit diesem Walkie-Talkie im exklusiven Marvel Spiderman | | 18.73 | Funktechnik & Zubehör (1,00) |
| 49 | 31001 - Walkie Talkie | Walkie-Talkie im exklusiven Hello Kitty Design. Kontaktiere Deine Freu: Hello Kitty | | 19.99 | Funktechnik & Zubehör (1,00) |

Automatische Duplikaterkennung

Integration aus verschiedenen Quellen

- Kinderwagen & Zubehör
- sonstige Baby- & Kinderartikel
- Film & Musik
- Büro & Schreibwaren
- Erotik
- Essen, Trinken & Tabakwaren
- Gesundheit & Wohlbefinden
- Geschenk
- Messe, Veranstaltung & Gastronomie

Details

Identifizier: 14978321

Name: HC-BM10 Drahtloses Babyfon weiß/blau

Beschreibung: Alarm beim Verlassen des Empfangsbereichs LED Lautstärkeanzeige am Receiver Überwachung Kinderzimmer Temperatur 2-Weg Kommunikation Nachtlicht

Hersteller: König

Preis: 65.95

Bild:

category: sonstige Baby- & Kinderartikel

Ausblick

- 2 Varianten verfügbar im **März 2012**
- **Community Edition**
 - Open Source, kostenlos
 - Einfache Matchstrategien
- **Software Suite** mit voller Funktionalität für Kategorisierung und Matching
 - für Business-Anwendungen