

Universität Leipzig
Institut für Informatik



**Matching und Integration großer
Produktkataloge**

Sabine Maßmann

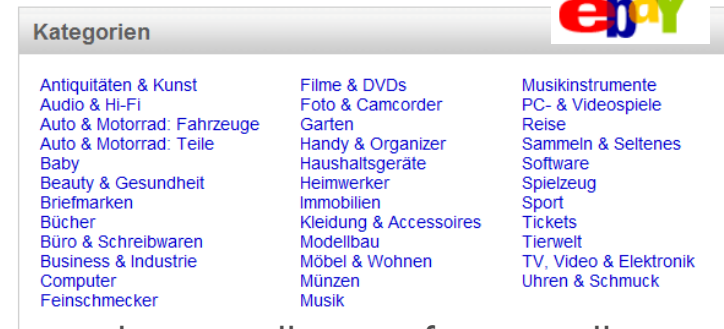
28.10.2010

Motivation

- Portale und Shops verwenden Produktkataloge
- Produktkatalog=
 - systematisch sortierte Sammlung von Produkt- oder Service-Information
 - Klassifikation erfolgt anhand eines gemeinsamen Merkmales, z.B. Hersteller, Feature, Farbe



- Portale
 - Benötigen Produktkatalog, der alle Produkte von den Händlern umfassen soll und somit idealerweise deren Produktkategorien beinhaltet
→ Erstellung eines allumfassenden Produktkatalogs
- Web-Shops
 - Wollen Produkte gerne bei verschiedenen Portalen einstellen
→ das Wissen über gleiche Kategorien erleichtert Einordnung von Produkten



Schwierigkeit

- Produktkataloge beinhalten hunderte bis tausende Kategorien und zehntausende bis millionen Produkte
- Kataloge können sehr verschieden sein
→ Heterogenität in Ausdrücken, Sprachen und Konzepten
- Kataloge verändern sich z.B. neue Trends → Anpassung nötig
- Beispiel

Ebay.de

TV, Video & Elektronik

- DVD-Player
- DVD-Recorder
- Videoprojektoren
- Leinwände

Musikinstrument

- Drums & Percussion
- Akustische Gitarre

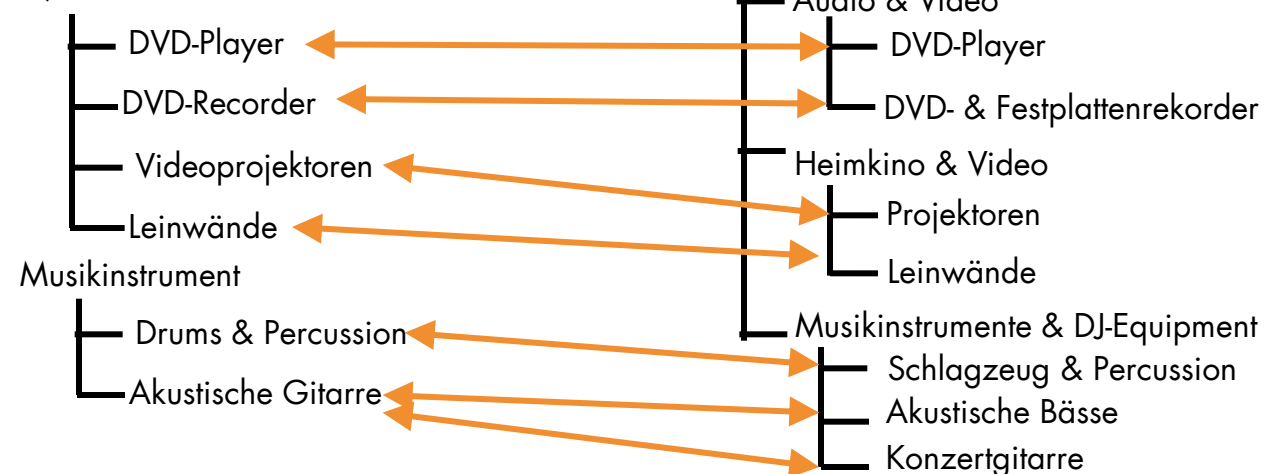
Amazon.de

Audio & Video

- DVD-Player
- DVD- & Festplattenrekorder
- Heimkino & Video
- Projektoren
- Leinwände

Musikinstrumente & DJ-Equipment

- Schlagzeug & Percussion
- Akustische Bässe
- Konzertgitarre



Anwendungen

- (Semi-)Automatische Erstellung eines integrierten Katalogs
- Einfache Anpassung bei Veränderung von Katalogen (Evolution)
- Nutzung der Korrespondenzen zwischen den Katalogen
 - Zum Auffinden zusätzlicher Produkte für eine Anfrage
 - Eine Anfrage über Produkte von mehreren Produktkatalogen
 - Automatische Eintragen von Produkten in die entsprechenden Kategorien verschiedener Anbieter

Unsere Lösung

1. Matching von Produktkatalogen
= Auffinden von Kategorien, die einander gleichen oder ähneln
 - *Eingabe*: 2 Kataloge (z.B. von Web-Shops oder Portalen)
 - *Ausgabe*: Korrespondenzen zwischen gleichen bzw. ähnlichen Kategorien

2. Integration von Produktkatalogen → *optional*
= Zusammenfügen der Kategorien unter Beachtung der Hierarchien und Beziehungen
 - *Eingabe*: 2 Kataloge und Korrespondenzen zwischen diesen Katalogen
 - *Ausgabe*: Integrierter Katalog

Automatisches Matching



- An Universität Leipzig wurde der Prototyp COMA entwickelt und erweitert zu COMA++
 - Generisches Matching-System
 - Ausnutzen von
 - Metadaten → Namen der Kategorien und evtl. Beschreibungen
 - Instanzen → Produktinformation wie z.B. Titel, Hersteller und Beschreibung
 - Strukturen → welche Kategorien sind Über- bzw. Unterkategorien
 - Externe Information z.B. Synonymlisten
 - Beliebiges Kombinieren der Ansätze
 - Strategien für das Matching großer Schemas und Ontologien

Do, H.H., E. Rahm: *COMA - A System for Flexible Combination of Schema Matching Approaches*. VLDB 2002

Aumüller D., H.-H. Do, S. Massmann, E. Rahm: *Schema and Ontology Matching with COMA++*. Sigmod 2005

Beispiel: Matching

Input:

- Ebay-Ausschnitt mit 40 Kategorien
- Amazon-Ausschnitt mit 43 Kategorien

Ergebnis:

- Korrespondenzen zwischen gleichen bzw. ähnlichen Kategorien

Nachbearbeitung:

- Überprüfung der Korrespondenzen

The screenshot shows the COMA++ software interface with a 'Repository Match Mapping View'. The interface is divided into several panes:

- Repository:** Lists two ontologies: 'amazon_Ausschnitt_owl (1)' and 'ebay_Ausschnitt_owl (1)'.
- Workspace:** Shows a 'Mapping1' operation with a table of details.
- Mapping1 Table:**

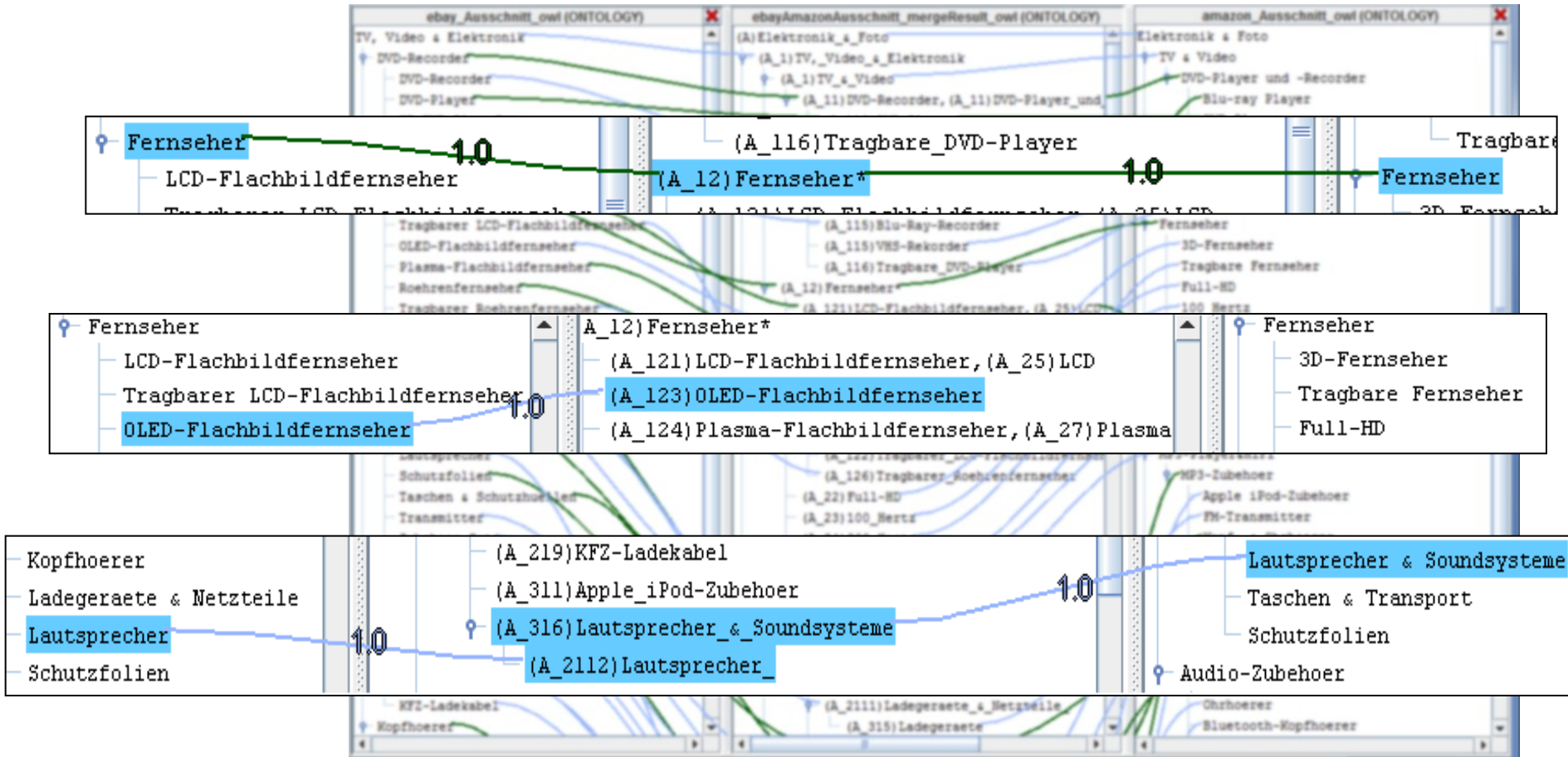
Name	Mapping1
Comment	COMA_OPT
Schemas	ebay_Ausschnitt_owl, a...
Operation	SCHEMA
Total	23 (0 + 23)
- Mapping View:** Displays two ontologies side-by-side with nodes connected by lines. The left ontology is 'ebay_Ausschnitt_owl (ONTOLOGY)' and the right is 'amazon_Ausschnitt_owl (ONTOLOGY)'. A color scale at the top indicates a match score from 1.0 (green) to 0.0 (red). The nodes are categorized into 'TV, Video & Elektronik', 'Fernseher', and 'Audio & Hi-Fi'. Lines connect specific nodes between the two ontologies, such as 'DVD-Recorder' to 'DVD-Recorder' and 'LCD-Flachbildfernseher' to 'Fernseher'.
- Source Node Information:** A table for the selected source node.
- Target Node Information:** A table for the selected target node.
- Search:** Input fields for searching nodes in both ontologies.

At the bottom of the window, it says 'Matching is done.'

Automatisches Integrieren

- Auch als *Mergen* bezeichnet
- Algorithmus verarbeitet die Korrespondenzen aus dem *Matching*
- Ziel ist, dass Kategorien, die :
 - in *beiden* Katalogen vorkommen, sollen im Ergebnis nur *eine* Kategorie darstellen
 - nur in *einem* der Katalogen vorkommen, sollen *eingefügt* werden - natürlich an der richtigen Stelle
 - nur *einem Teil* abgedeckt sind, sollen entsprechend in der Hierarchie *eingefügt* werden
- Eingabe:
 - 2 Kataloge und Korrespondenzen zwischen deren Kategorien
- Ausgabe:
 - 1 integrierter Katalog und 2 Mappings bestehend aus den Korrespondenzen der gegebenen Kataloge zum integrierten Katalog

Beispiel: Integration



Kategorien

Ebay-Ausschnitt
40

Integriert
68

Amazon-Ausschnitt
43

Vorteile

- Matching mit COMA++
 - (semi-)automatisches Matching
 - Strategien für große Kataloge
 - Evaluation anhand verschiedener Szenarien
- Integrations-Algorithmus
 - Automatisch und optimiert für große Datensets:
Beispiel: 2 Versionen vom Ebay-Katalog
 - Kataloge mit 24480/ 22513 Kategorien, 23805 Korrespondenzen
 - Zeit zum Laden und Speichern 5 s
 - Zeit zur Ausführung des Merge-Algorithmus 11 s
 - Ergebnis: Katalog mit 23123 Kategorien
 - Neben dem integrierten Katalog werden auch die Korrespondenzen zwischen Ausgangskatalogen und integriertem Katalog erzeugt
 - man kann zurückverfolgen, woher die Kategorien stammen
 - Weiternutzung zur Anfrage-Abarbeitung möglich

Session 1: Informationsqualität von Web-Shops

- Kategorisierung von Produktangeboten in großen Katalogen
Hanna Köpcke
- Dublettenerkennung und Bereinigung von Webdaten
Hanna Köpcke
- Matching und Integration großer Produktkataloge
Sabine Maßmann