

Verfahren zur Dublettenerkennung und - behandlung

Kolloquium Wissensinfrastruktur, 24. Juni 2016

Jochen Schirrwagen, Universitätsbibliothek Bielefeld

Agenda

- Dubletten
- Dublettenerkennungsmethoden
- Dublettenfusion

Begriffs-Definition

- Duplikate: verschiedene Datensätze, die dasselbe Realweltobjekt (Entität) repräsentieren
- Dublikaterkennung: Identifizierung von Dubletten (record linkage, object identification, entity resolution)

Auftreten

- Bei Mehrfacheingaben eines Datensatzes in ein System (Katalog, Repository, ...)
- Bei Informationsintegration
 - Zusammenführung von Objekten aus unterschiedlichen
 Systemen (Data-Warehouse-System, Metadatenaggregator, ...)
- Weitere?

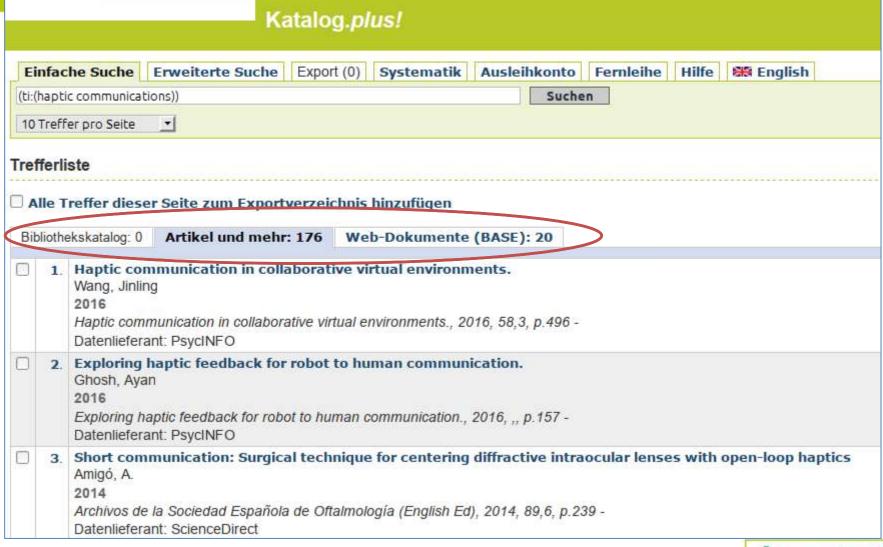
Fehlertypologie in Datensätzen

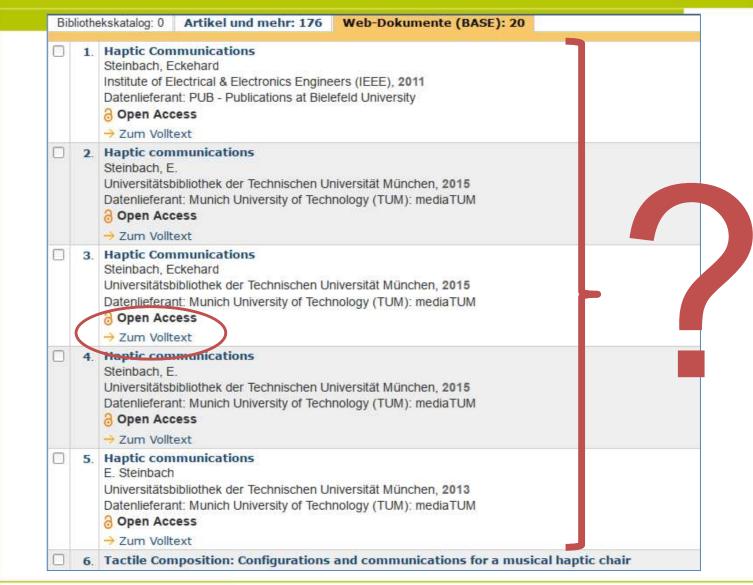
- Semantische Fehler, z.B. Altdaten mit Feldern und Werten, die aktuellem Datenmodell nicht mehr genügen
- Syntaktische Fehler, z.B. Werte in falschem Feld oder in falscher Reihenfolge ("Peter Müller" statt "Müller, Peter")
- Lexikalische Fehler ("Müller" statt "Mueller")
- Typographische Fehler ("Unsin")
- Kodierungsfehler (z.B. Zeichenkette nicht in UTF-8)

Bedeutung

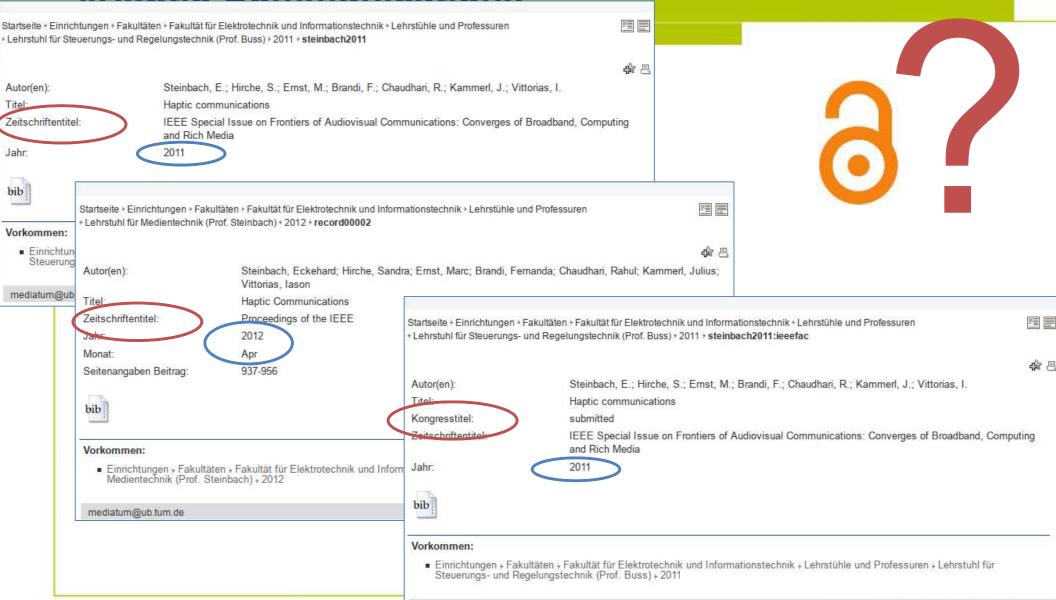
- Qualitätsaspekt
 - Z.B. möglichst genaue, eindeutige Trefferlisten
 - Voraussetzung für verlässliche Mehrwertdienste wie:
 - (Nutzungs-) Statistiken
 - "Claiming"-Dienste (Autor <-> Publikationen)
- Informationsanreicherung und –korrektur
- Weitere Ideen?





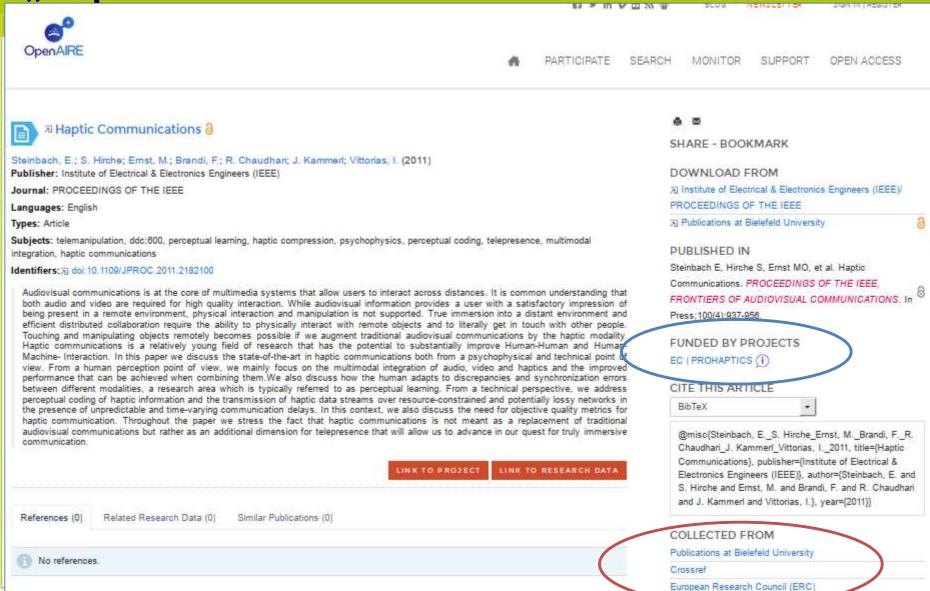






mediatum@ub.tum.de

Datenschutz Impressum 23.06.2016 20:03:26



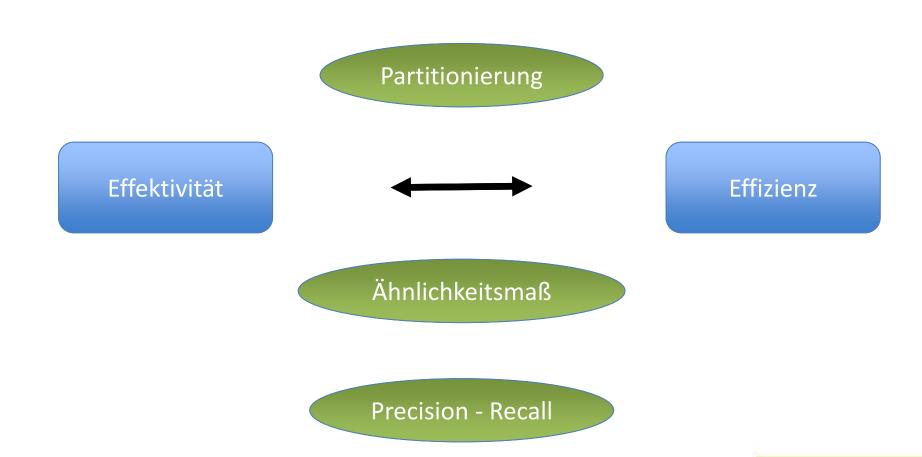
d.de

PUB OAI Metadaten (keine Projektinformation enthalten)

```
version="1.0" encoding="UTF-8"?>
  code xmlns="http://www.openarchives.org/OAI/2.0/oai_dc/"
       xmlns:oai dc="http://www.openarchives.org/OAI/2.0/oai dc/"
       xmlns:dc="http://purl.org/dc/elements/1.1/"
       xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
       xsi:schemalocation="http://www.openarchives.org/OAI/2.0/oai_dc/ http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
    fills>Haptic Communications</dc titls>
          r>Steinbach, Eckehard</d
          r>Hirche, Sandra</do
          r>Ernst, Marc 0.</d
          >Brandi, Fernanda</
          >Chaudhari, Rahul</
          >Kammerl, Julius</
          >Vittorias, Iason</
          t>multimodal integration</d
          t>psychophysics</d
          t>perceptual coding</d
          >haptic compression</
          t>telemanipulation</doi
</pre>
          t>haptic communications</dr subtect>
          t>telepresence</dd:subj
          t>perceptual learning</dc:subject>
          t>ddc:600</di
            >Institute of Electrical & Electronics Engineers (IEEE)</dc:publishes>
        >info:eu-repo/semantics/article</ductype>
        >doc-type:article</dc:type>
         ifler>https://pub.uni-bielefeld.de/publication/2422811</dc:identi
         ifler>https://pub.uni-bielefeld.de/download/2422811/2422812</dc:identifle>
         :e>Steinbach E, Hirche S, Ernst MO, et al. Haptic Communications. <en&gt;PROCEEDINGS OF THE IEEE, FRONTIERS OF AUDIOVISUAL COMMUNICATIONS&lt;/en&gt;. In Press;100(4):937-956.</dc>
          on>info:eu-repo/semantics/altIdentifier/doi/10.1109/JPROC.2011.2182180</doi/10.1109/JPROC.2011.2182180
           info:eu-repo/semantics/altIdentifier/issn/0018-9219</d>
    relition>info:eu-repo/semantics/altIdentifier/issn/1558-2256</
<dc:relation>info:eu-repo/semantics/altIdentifier/wos/000302088600012</dc:relation>
         s>info:eu-repo/semantics/openAccess</de:rights>
```

Google	Haptic Communications	Scholar	7 Ergebnisse (0,01 Sek.)
cholar	Ungefähr 58.800 Ergebnisse (0,06 Sek.)	Alle Versionen	Haptic communications <u>E Steinbach</u> , S Hirche, <u>M Ernst</u> , F Brandi Proceedings of the, 2012 - ieeexplore ieee.org ABSTRACT Audiovisual communications is at the core of multimedia systems that allow users to interact across distances. It is common understanding that both audio and video are required for high-quality interaction. While audiovisual information provides a user with a
eliebige Zeit eit 2016	Tipp: Suchen Sie nur nach Ergebnissen auf Deutsch. Sie können Ihre Sprache in den Schola Haptic communications E Steinbach, S Hirche, M Ernst, F Brandi Proceedings of the, 2012 - ieeexplore.ieee.or ABSTRACT Audiovisual communications is at the core of multimedia systems that allow users to interact across distances. It is common understanding that both audio and video are required for high-quality interaction. While audiovisual information provides a user with a Zitiert von: 39 Ähnliche Artikel Alle 8 Versionen Zitieren Speichern	DI .	Titler von: 39 Ahnliche Artikel Zitieren Speichem Haptic Communications E Steinbach, S Hirche, M Ernst, F Brandi Proceedings of the, 2012 - infons.pl Audiovisual communications is at the core of multimedia systems that allow users to interact across distances. It is common understanding that both audio and video are required for high-quality interaction. While audiovisual information provides a user with a satisfactory Zitieren Haptic Communications E STEINBACH, S HIRCHE, M ERNST, F BRANDI Proceedings of the, 2012 - cat.inist.fr Résumé/Abstract Audiovisual communications is at the core of multimedia systems that allow users to interact across distances. It is common understanding that both audio and video are required for high-quality interaction. While audiovisual information provides a Zitieren Haptic Communication IE Steinbach - nano.ei.tum.de Abstract: In this talk I will discuss selected aspects concerning the acquisition, coding, transmission, display and perception of haptic information. In the first part, the focus will be on the communication of kinesthetic signals for teleoperation. The second part will focus Zitieren pranon The What, Why, and How of Haptic Communication E Steinbach - Imt.ei.tum.de A System for High Precision Glass-to-Glass Delay Measurements in Video CommunicationIEEE International Conference on Image Processing (ICIP 2016), Phoenix, AZ, USA, September 2016 Energy Prediction for Teleoperation Systems that Combine the Time Domain Zitieren pranon The What, Why, and How of Haptic Communication E Steinbach - Imt.ei.tum.de A System for High Precision Glass-to-Glass Delay Measurements in Video Communication Esteinbach - Imt.ei.tum.de A System for high Precision Glass-to-Glass Delay Measurements in Video Communication Esteinbach - Imt.ei.tum.de A System for high Precision Glass-to-Glass Delay Measurements in Video Communication Esteinbach - Imt.ei.tum.de A System for high Precision Glass-to-Glass Delay Measurements in Video Communication In Estein

Zielkonflikt bei der Umsetzung von Verfahren zur Dublettenerkennung



→ www.uni-bielefeld.de

Effektivität

- Gute Ergebnisse, also Duplikate möglichst genau und vollständig erkennen
- Mit Hilfe von Ähnlichkeitsmaßen (verwandt mit Distanzmetriken)
- i.d.R. Vergleich von Zeichenketten, alternativ Vergleich strukturierter Datensätze etc.
- D.h. Maßzahl ermitteln, z.B. im Bereich {0..1}
- Schwellwert festlegen
- Im Idealfall: Maßzahl = 1, also zwei Datensätze sind identisch
- Beurteilung der Güte der gewählten Methode mit den Maßen
 - <u>Precision</u> (Präzision, Genauigkeit): Verhältnis gefundener relevanter Dokumente zur Zahl aller gefundenen Dokumente
 - Recall (Vollständigkeit): Verhältnis zwischen gefundenen relevanten Dokumenten und Gesamtzahl aller vorhandenen relevanten Dokumente



Ähnlichkeitsmaße (Auswahl)

- Edit-Distanz:
 - z.B. Levenshtein-Distanz (minimale Anzahl an Edit-Operationen, um Zeichenkette A in Zeichenkette B zu überführen)
 - Tolerant gegenüber lexikalische / typograph. Fehler
 - Schlecht bei syntakt. Fehlern (Feldvertauschungen)
- Jaro-Winkler-Ähnlichkeit: geeignet für kurze Wörter, berücksichtigt Transposition einzelner Buchstaben

Kosinus-Ähnlichkeit

• Betrachtung von Zeichenketten als Vektoren im (eukl.) Vektorraum

•
$$\cos \theta = \frac{a*b}{\|a\|*\|b\|}$$
 b

- Ähnliche Zeichenketten liegen als Vektoren dicht beieinander -> kleiner Winkel
- Tolerant gegenüber Position u. Reihenfolge einzelner Terme in einer Zeichenkette ("Peter Müller" entspricht "Müller, Peter")
- Aber schlecht bei typogr., phonet. Fehlern

tf-idf-Maß (Termgewichtung)

- Wichtung von Termen, so dass Terme mit höherem Gewicht Kosinus-Ähnlichkeit stärker beeinflussen als Terme mit niedrigerem Gewicht
- tf = term frequency (Vorkommenshäufigkeit)
 - Lokaler Gewichtungsfaktor, d.h. Häufigkeit eines Terms in einer Zeichenkette
- idf = inverse document frequency (inverse Dokumenthäufigkeit)
 - Globaler Gewichtungsfaktor, d.h. Häufigkeit eines Terms im Dokumentenkorpus insgesamt

Effizienz

- Wie gut skaliert die Menge zu vergleichender Datensätze mit der Laufzeit?
- Paarweiser Vergleich jedes Datensatzes mit jedem anderen heißt:
 - $\frac{n^2-n}{2}$ Vergleiche, aber exponentielle Laufzeit $\mathcal{O}(n^2)$
- Optimierungsansatz: Partitionierung der Datensätze
 - Reduzierung der Laufzeit auf $O(n * \log n)$



Partitionierung

- Blocking / Clustering
 - Unterteilung "grob ähnlicher" Datensätze in Partitionen
 - Z.B. anhand bestimmter Attribute, wie das Publikationsjahr
 - Ggf. überlappende Partitionen
 - Schwierigkeit besteht im Finden guter Blocking-Strategie
 - Partition soll tatsächliche Duplicate enthalten
- "Fenster-Technik" (sliding window, sorted neighborhood)
 - Sortierung der Datensätze nach bestimmten Schlüssel / Eigenschaften
 - Mögliche Duplikate liegen nahe beieinander
 - Vergleich aller Datensätze, die in einem Fenster der Größe n liegen
- In der Praxis Kombination beider Verfahren

Datenfusion (Merging)

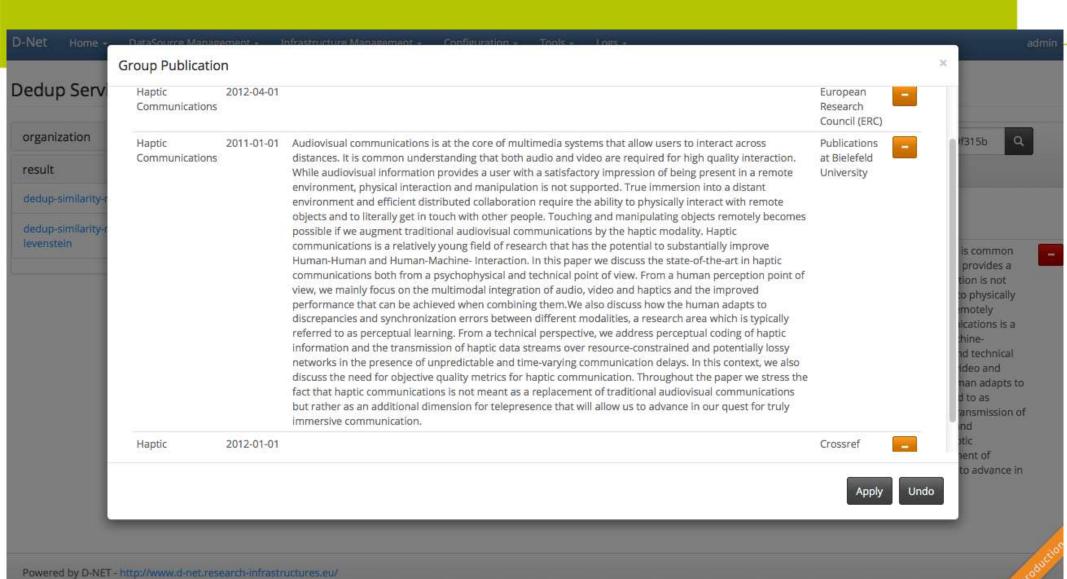
- Entweder
 - Wahl eines repräsentativen Datensatzes
 - Aber nach welchen Kriterien?
 - Was passiert mit den Dubletten?
- Oder
 - Erzeugung eines neuen Datensatzes
 - Aus den Werten der Dubletten
 - Im Ergebnis ein angereicherter Datensatz
 - Wie mit evtl. vorhandener fehlerhafter Information in Dubletten umgehen?

Dubletten-Kuration in OpenAIRE

- In Ergänzung zu automatischen Verfahren zur Zeit auf Ebene von
 - Organisationen
 - Bibliographischen Metadaten
- Erlaubt das einfache Gruppieren von Datensätzen (hinzufügen / entfernen)



Dubletten-Kuration in OpenAIRE





Quellen

- Krause, Th. (2012). Entwurf und Implementierung einer effizienten
 Dublettenerkennung für große Adressbestände. Masterarbeit,
 http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:hbz:832-epub-3667
- Atzori, Cl. (2016). gDup: an integrated and scalable graph deduplication system. Dissertation, https://etd.adm.unipi.it/t/etd-05092016-090250/
- Naumann, F. (2007). Methoden der Dublettenerkennung.
 https://hpi.de/fileadmin/user_upload/fachgebiete/naumann/publications/2007/40-43_IQR_Dubletten_erkennen.pdf
- Wendykier, P. (2013). Deduplication of metadata harvested from Open Archives Initiative repositories. Information Services & Use, 33(2), 77-86. DOI: 10.3233/ISU-130695