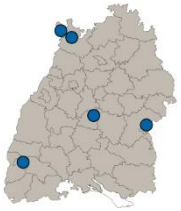
NETZWERK SELTENE ERKRANKUNGEN BADEN-WÜRTTEMBERG KOMPETENZZENTRUM



ulm university universität ulm

Hochschule Ulm
University of Applied Sciences

Nutzung von String-Ähnlichkeitsmaßen in Talend Open Studio zur Desambiguierung von Autorennamen aus PubMed

13. DVMD-Fachtagung 12.03.2015

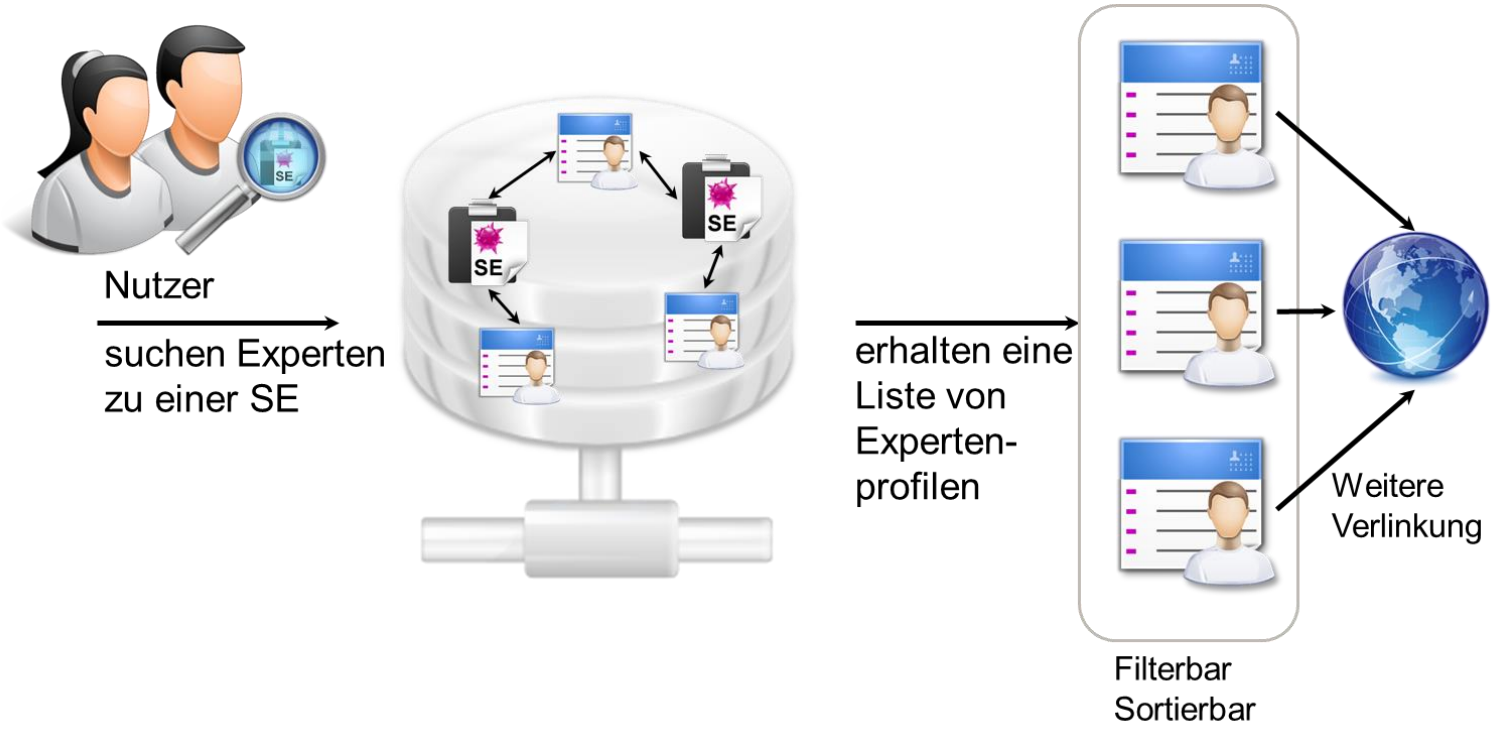
Andreas Pflugrad, Jochen Bernauer
Institut für Informatik, Hochschule Ulm

DVMD

Hintergrund – Aufbau einer Profildatenbank

Ziel: Unterstützung von Patienten, Ärzten und Angehörigen

bei der Suche nach Experten für Seltene Erkrankungen (SE)



Hintergrund – Generierung von Expertenprofilen

1. Extraktion von Literatur-Metadaten

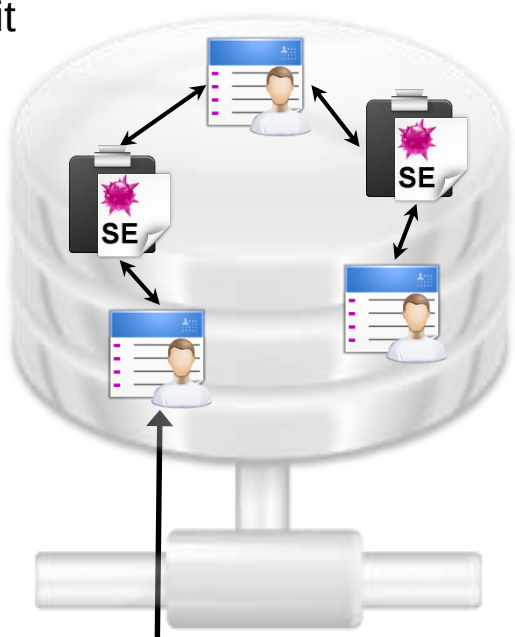
2. Berechnung von Indikatoren und Generierung von Profilen

3. Verknüpfung von Profilen mit Seltenen Erkrankungen

(Beispiel)

Autor + Metadaten

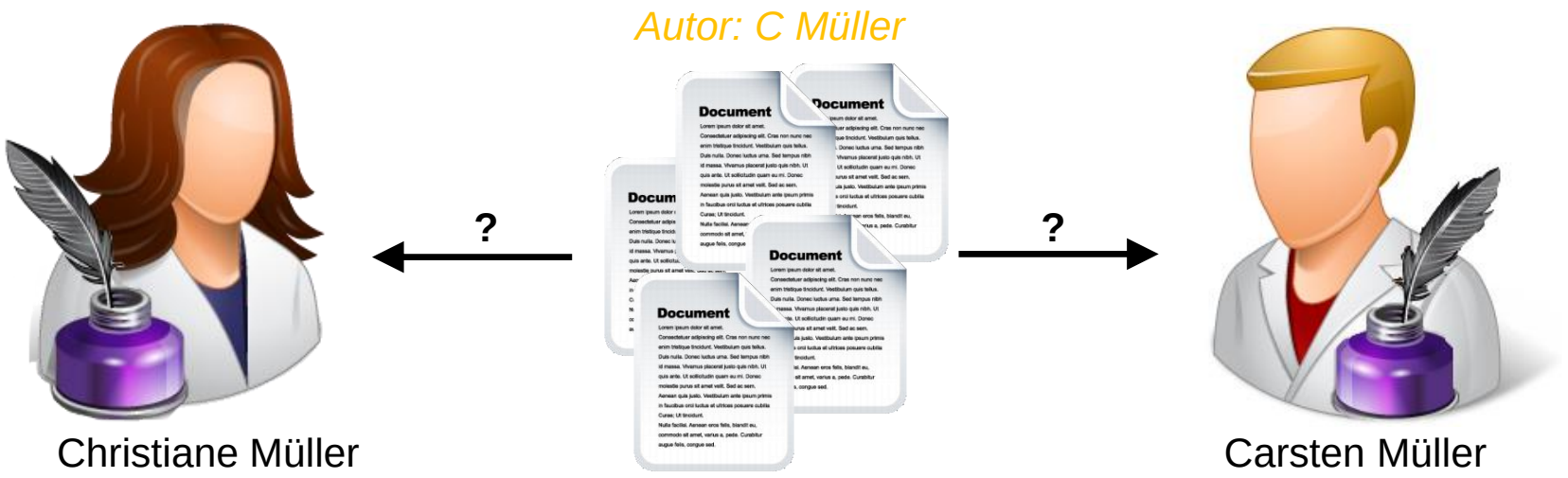
Experten-profil



- Publikationsart
- Journal
- Autorpositionen
- Institutionen
- Schlüsselbegriffe

4. Anreicherung der Profile mit Daten aus dem Web

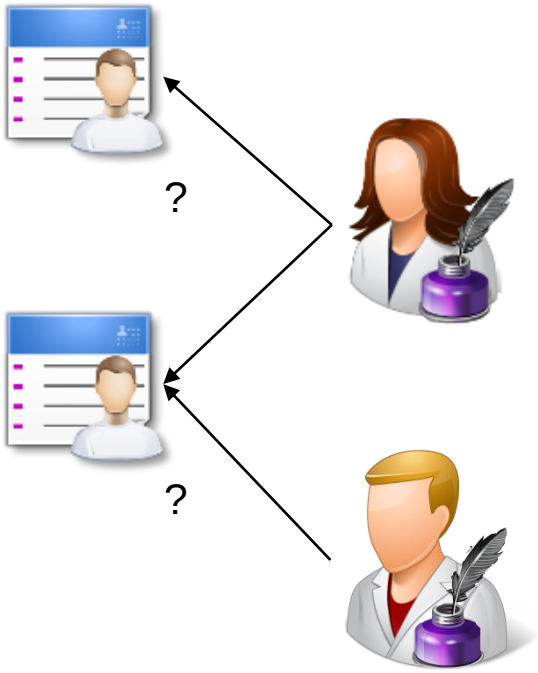
Hintergrund – Problematik der Namensunterscheidung



- Desambiguierung: zuordnen der einzelnen Namen (Publikationen) zu den richtigen Personen anhand der Metadaten jeder Publikation

Hintergrund – Problematik der Namensunterscheidung

- Problem: hoher Rechenaufwand
  - Paarweiser Vergleich aller Publikationen nicht realistisch
- Lösung: Bildung von Gruppen
  - Gleiche/Ähnliche Publikationen werden gruppiert
  - Paarweiser Vergleich nur innerhalb von Gruppen
  - Kriterium für die Gruppierung: Autorennamen



Schritt 1: Bildung von Gruppen → Schritt 2: Desambiguierung innerhalb der Gruppen

Hintergrund – Bildung von Gruppen

- Ansatz 1: anhand exakter Namensübereinstimmung
  - Kleinere Gruppen
  - Keine falsch-positiven Übereinstimmungen
  - Potentiell vermehrt falsch-negative Ergebnisse

C Müller = C Müller

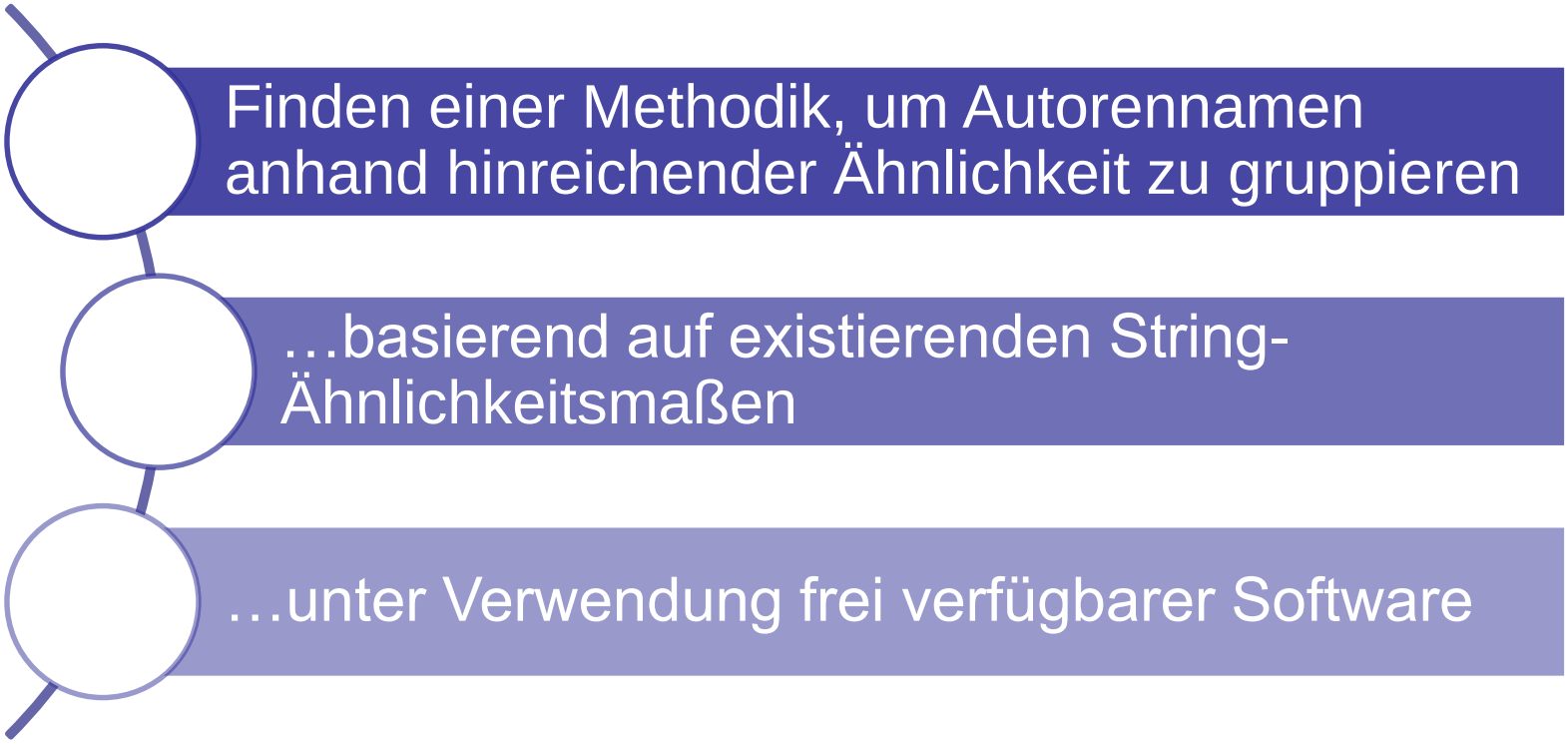
- Ansatz 2: anhand von Namensähnlichkeiten
  - Größere Gruppen
  - Mehr falsch-positive Übereinstimmungen
  - Ermöglicht die Berücksichtigung von Tippfehlern, Doppelnamen etc.

C Müller = C Muller

F Lehmann = F Lehman

E Koch = E Koch-Heinze

Zielstellung der vorliegenden Untersuchung



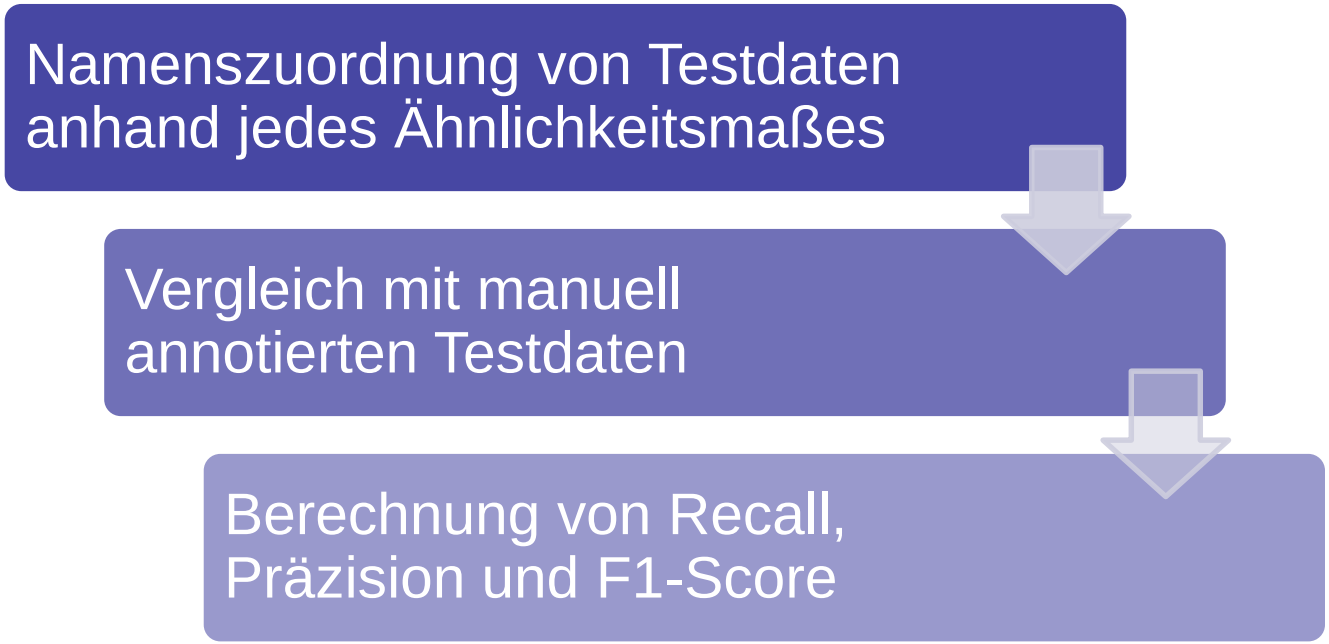
Methoden – manuelle Annotation von Testdaten

- Zufällige Auswahl von vorläufigen Namenszuordnungen (aus einem intuitiven Verfahren)
- Manuelle Validierung der Namenszuordnung

Name	Zugeordneter Name	Zugehörig
LALLIER	LAULIER	1
LALLIER	LOLLIER	0
LAMBRECHTS	LAMBRECHT	1
LAMBRECHTS	LEMBRECHTS	0
LAMOTTE-BARRILLON	LAMOTTE BARILLON	1

- Kategorisierung von Namensunterschieden
  - Änderungsart: Austausch/Hinzukommen/Wegfallen v. Zeichen
  - Anzahl betroffener Zeichen: ein Zeichen, mehrere Zeichen
  - Auswirkung: Zuordnung/keine Zuordnung

Methoden – Vergleich der Ähnlichkeitsmaße



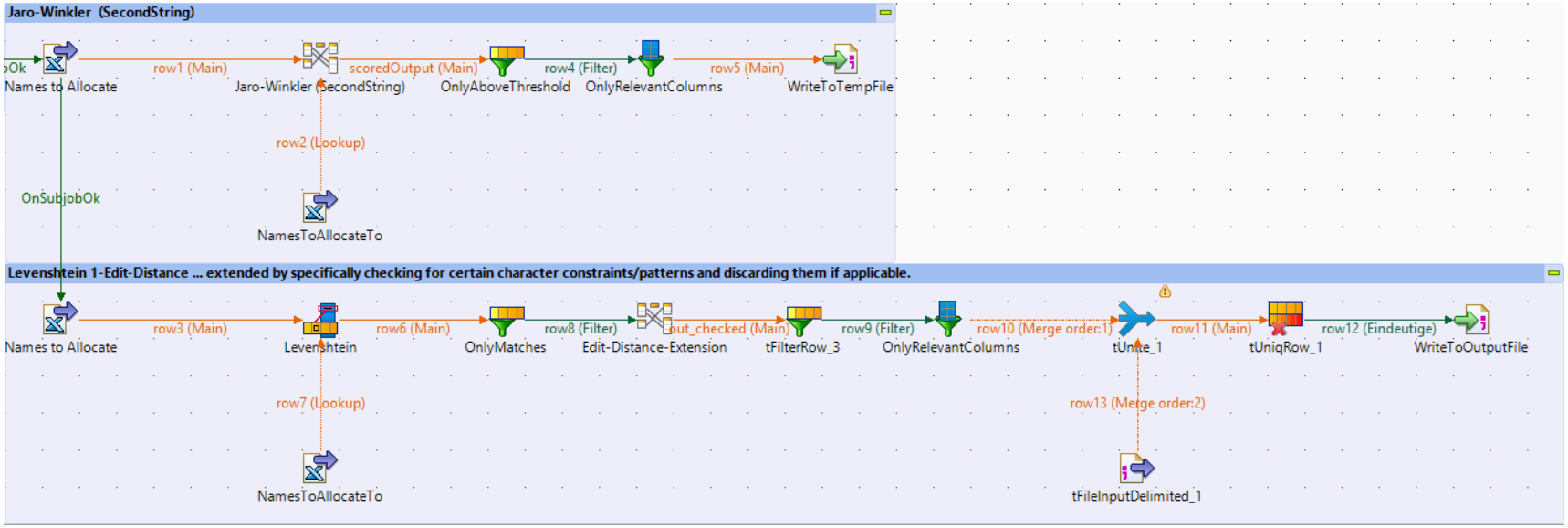
- Für jedes Ähnlichkeitsmaß einzeln
- Für vielversprechende Maße in Kombination

Methoden – verwendete Ähnlichkeitsmaße

- Basiert auf Edit-Distanz
  - Levenshtein Editierdistanz
  - Jaro-Winkler
  - Monge-Elkan
- Phonetisch
  - Metaphone / Double Metaphone
- Tokenbasiert
  - Jaccard

Methoden – verwendete Software

- Talend Open Studio (TOS)
  - Freie komponentenbasierte Softwaresuite zur Datenintegration



- SecondString Bibliothek
  - Sammlung von String-Ähnlichkeitsmaßen, Java-Klassenbibliothek

Ergebnisse

- Einzelne Ähnlickeitsmaße

Ähnlichkeitsmaß	Recall	Präzision	F1-Score
Levenshtein (1-Edit-Distance)	0,800	0,727	0,762
Metaphone	0,925	0,673	0,779
Double-Metaphone	0,831	0,618	0,708
Jaro-Winkler	0,846	0,753	0,797
Jaccard	0,258	0,470	0,334
Monge-Elkan	0,557	0,407	0,470

- Kombinationen

Jaro-Winkler ∩ Levenshtein	0,749	0,785	0,766
Jaro-Winkler ∪ Levenshtein	0,943	0,685	0,794
Jaro-Winkler ∩ Metaphone	0,875	0,774	0,821
Jaro-Winkler ∪ Metaphone	0,980	0,650	0,782
Levenshtein ∩ Metaphone	0,744	0,757	0,750
Levenshtein ∪ Metaphone	0,980	0,657	0,787

Erweiterung des Levenshtein Ähnlichkeitsmaßes

- Erweiterung der Levenshtein Editierdistanz (1 Zeichen)
- Ablehnung einer Namenszuordnung, wenn der Zeichenunterschied bestimmte Kriterien erfüllt.
- Kriterien basieren auf erkannten Systematiken bei Zeichenunterschieden in den Testdaten

Wang ≠ Wong

- Testergebnisse mit Erweiterung:

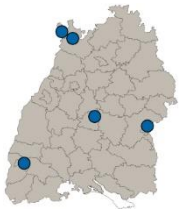
Ähnlichkeitsmaß	Recall	Präzision	F1-Score
Levenshtein (E)	0.642 **(-0.158)**	0.917 **(+0.19)**	0.755 **(-0.007)**
Jaro-Winkler ∪ Levenshtein (E)	0.958 **(+0.015)**	0.682 **(-0.003)**	0.797 **(+0.003)**
Levenshtein (E) ∪ Metaphone	0.971 **(-0.009)**	0.672 **(+0.015)**	0.794 **(+0.007)**

Fazit

- Umsetzung von Methoden zur Namenszuordnung in TOS gut machbar
- Verbesserte Möglichkeiten mit Hinzunahme externer Bibliotheken und eigenen Modulen
- Ähnlichkeit von Nachnamen = relativ spezifisches Problem, jedoch vielfältig für die Desambiguierung von Personen anwendbar
- Internationalität der untersuchten Methoden noch zu verifizieren
- F1 score als Vergleichsmaß für String-Ähnlichkeit wird in der Literatur hinterfragt

Vielen Dank für Ihre Aufmerksamkeit!

NETZWERK SELTENE ERKRANKUNGEN BADEN-WÜRTTEMBERG KOMPETENZZENTRUM



ulm university universität ulm

Hochschule Ulm University of Applied Sciences

Kontakt: Andreas Pflugrad
Institut für Informatik
Hochschule Ulm

pflugrad@hs-ulm.de

www.seltene-erkrankungen.info/Expertensuche

Referenzen

- Torvik VI, Smalheiser NR. *Author Name Disambiguation in MEDLINE*. ACM Trans Knowl Discov Data 2009; 3(3).
- Pflugrad A, Jurkat-Rott K, Lehmann-Horn F, Bernauer J. *Towards the automated generation of expert profiles for rare diseases through bibliometric analysis*. Studies in health technology and informatics 2014; 198:47–54.
- Liu W, Islamaj Doğan R, Kim S, Comeau DC, Kim W, Yeganova L et al. *Author name disambiguation for PubMed*. J Assn Inf Sci Tec 2014; 65(4):765–81.
- Liu J, Lei KH, Liu JY, Wang C, Han J. *Ranking-based name matching for author disambiguation in bibliographic data*. In: the 2013 KDD Cup 2013 Workshop. p. 1–8 .
- Da Silva R, Stasiu R, Orengo VM, Heuser CA. *Measuring quality of similarity functions in approximate data matching*. Journal of Informetrics 2007.