Automatische Klassifikation elektronischer Dokumente

Mathias Lösch

Universitätsbibliothek Bielefeld

Mathias.Loesch@uni-bielefeld.de

Kolloquium Wissensinfrastruktur, 20. Mai 2011

Hintergrund

- DFG-Projekt »Automatische Anreicherung von OAI-Metadaten«
- Förderung Oktober 2009–September 2011
- Partner:
  - Universitätsbibliothek Bielefeld
  - Abteilung für geisteswissenschaftliche Fachinformatik, Unversität Frankfurt/Main
  - Institut für automatische Sprachverarbeitung, Universität Leipzig

Agenda

1. Motivation
2. Automatische Klassifikation
3. Use Cases
4. Zusammenfassung

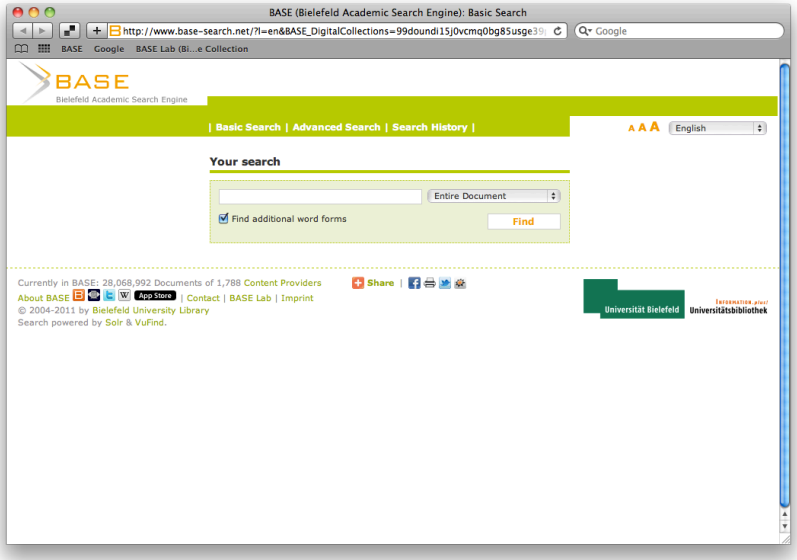
1 **Motivation**

2 Automatische Klassifikation

3 Use Cases

4 Zusammenfassung

Motivation



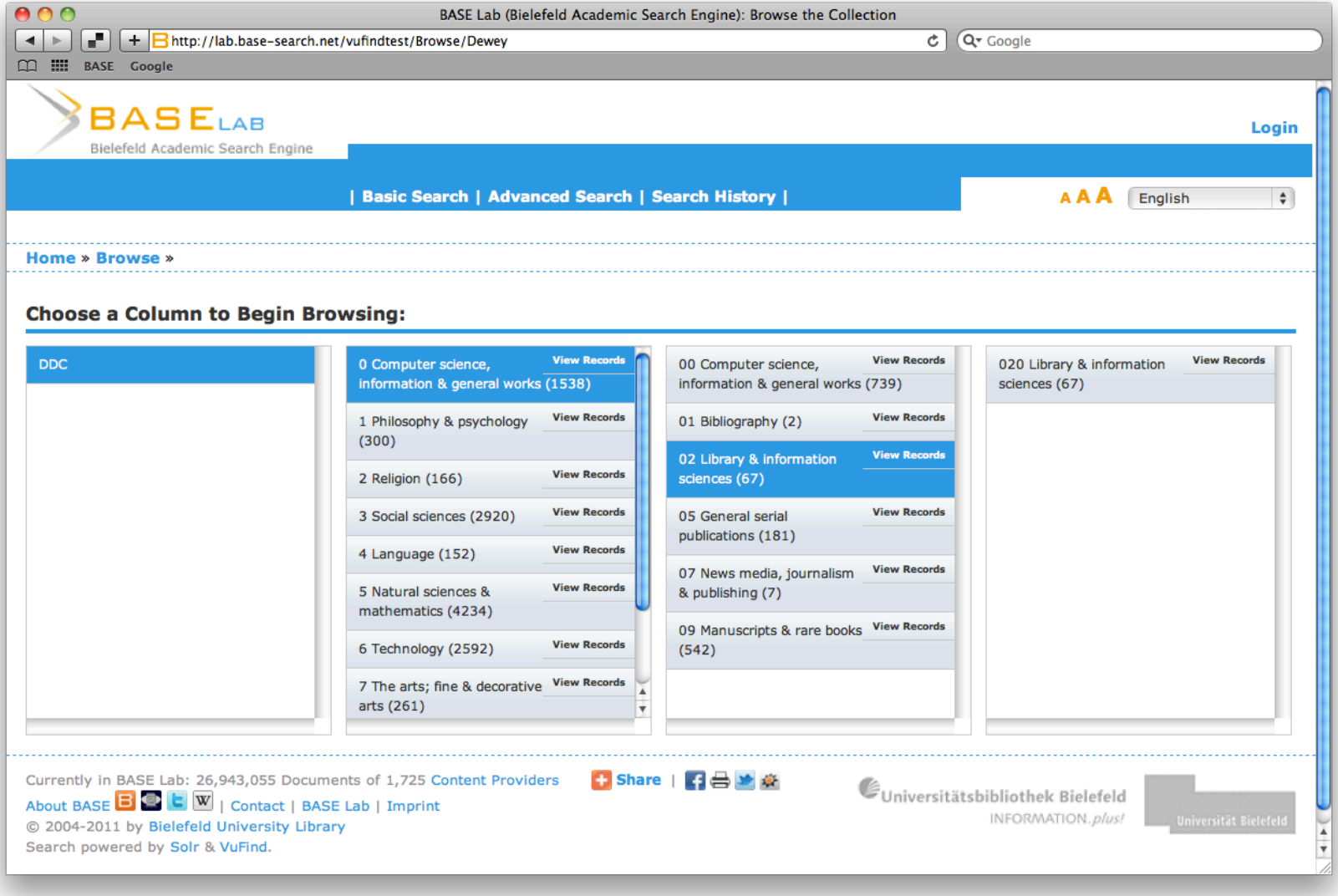
BASE
Bielefeld Academic Search Engine

- Wissenschaftlicher Suchservice
- Zugriff auf > 28 Mio Dokumente
- Aggregation der Inhalte von > 1.700 Dokumentenserver

Universität Bielefeld
Information.plus! Universitätsbibliothek
DFG

Motivation

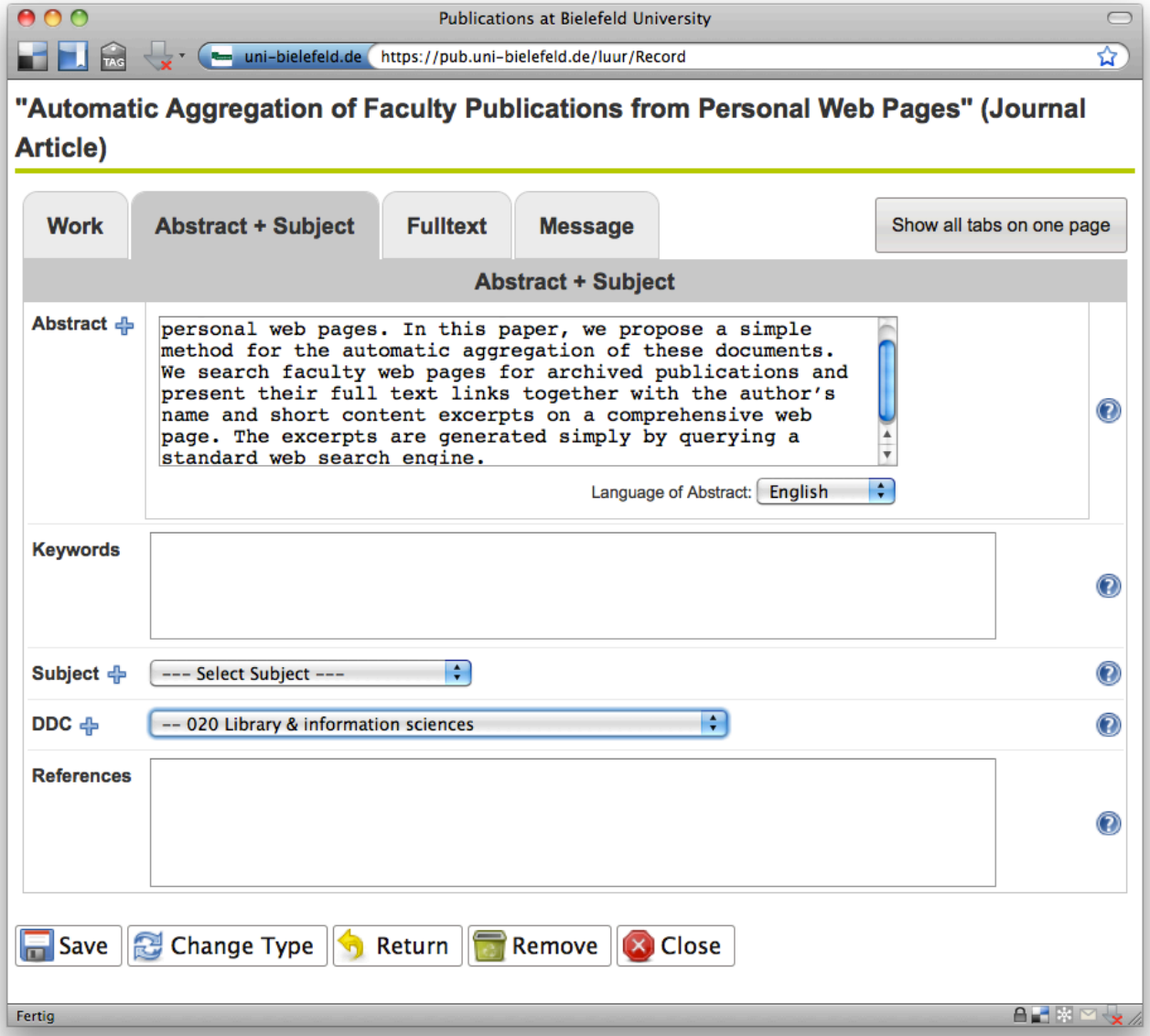
BASE Browsing



Motivation

- Trend zu disziplinspezifischen Dokumentenservern
  - arXiv.org (Physik)
  - PubMed (Life sciences)
  - Econstor (Wirtschaft)
  - SSOAR (Sozialwissenschaft)
  - ...
- Interesse an automatischer Extrahierung relevanter Subsets aus der BASE-Datenbasis

Motivation


Publications at Bielefeld University
uni-bielefeld.de
https://pub.uni-bielefeld.de/luur/Record
"Automatic Aggregation of Faculty Publications from Personal Web Pages" (Journal Article)
Work
Abstract + Subject
Fulltext
Message
Show all tabs on one page
Abstract + Subject
Abstract
personal web pages. In this paper, we propose a simple method for the automatic aggregation of these documents. We search faculty web pages for archived publications and present their full text links together with the author's name and short content excerpts on a comprehensive web page. The excerpts are generated simply by querying a standard web search engine.
Language of Abstract: English
Keywords
Subject
--- Select Subject ---
DDC
-- 020 Library & information sciences
References
Save
Change Type
Return
Remove
Close
Fertig

1 Motivation

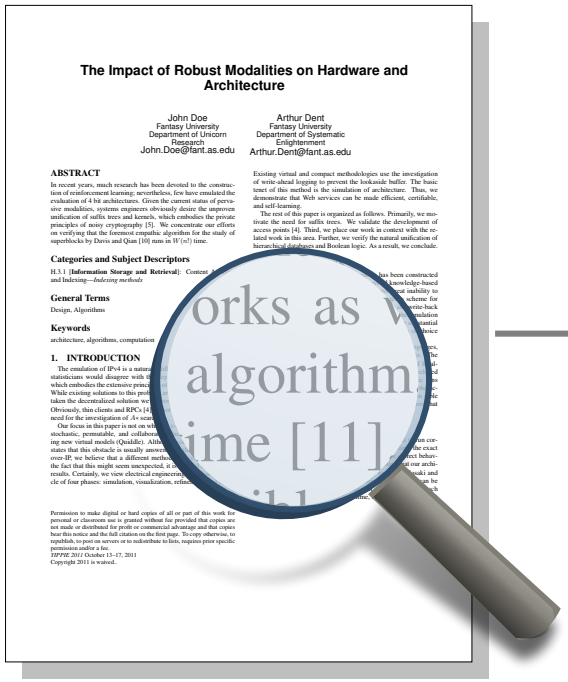
2 **Automatische Klassifikation**

3 Use Cases

4 Zusammenfassung

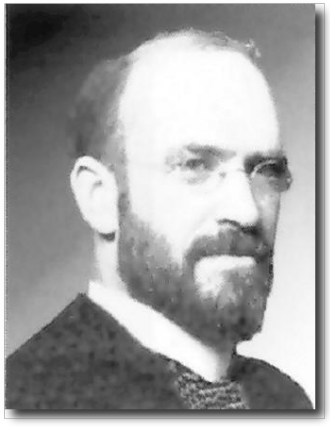
Lösungsansatz

Automatische Klassifikation von Dokumenten auf Basis ihres Inhalts

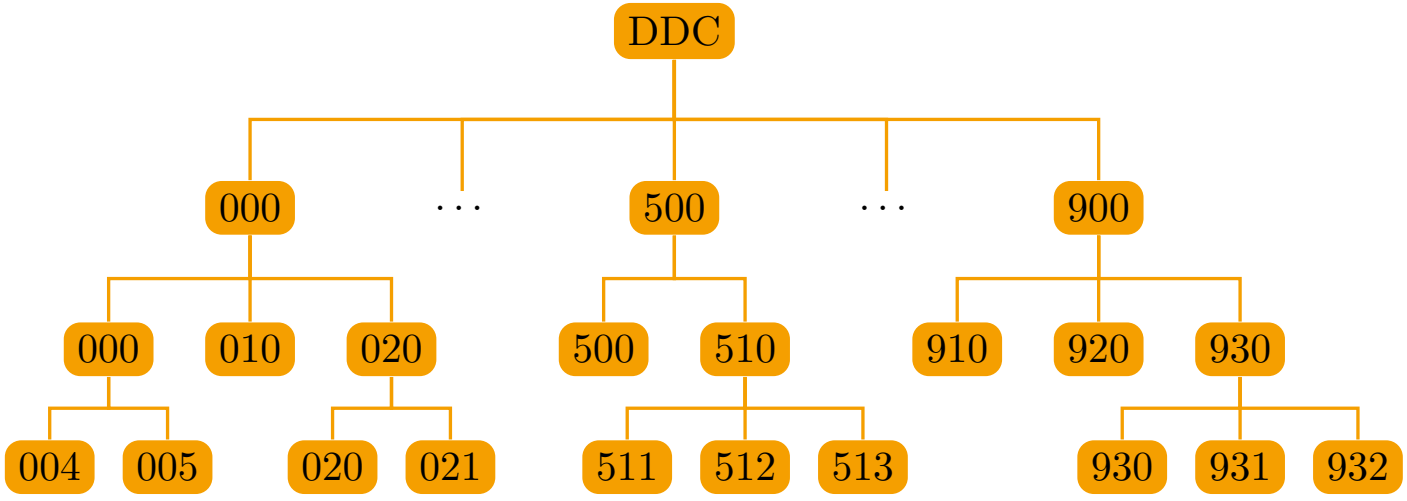


INFORMATIK

Dewey Decimal Classification



Melvil Dewey
(1851–1931)



Dewey Decimal Classification

Vorteile der DDC

- universal
- international starke Verbreitung (~200.000 Bibliotheken weltweit)
- Hierarchische Baumstruktur: maschinell einfach traversierbar
- Numerische Notation: Sprachunabhängige Kodierung der Klassen
- Dezimalstruktur: auf-/absteigende Traversierung durch Trunkierung/Expansion der Nummern einfach möglich
- Durch Empfehlung von DINI in der deutschen Repository-Landschaft meist-verwendete Klassifikation

»Das Lehren der Sprache ist hier kein Erklären, sondern ein Abrichten.« (L. Wittgenstein)

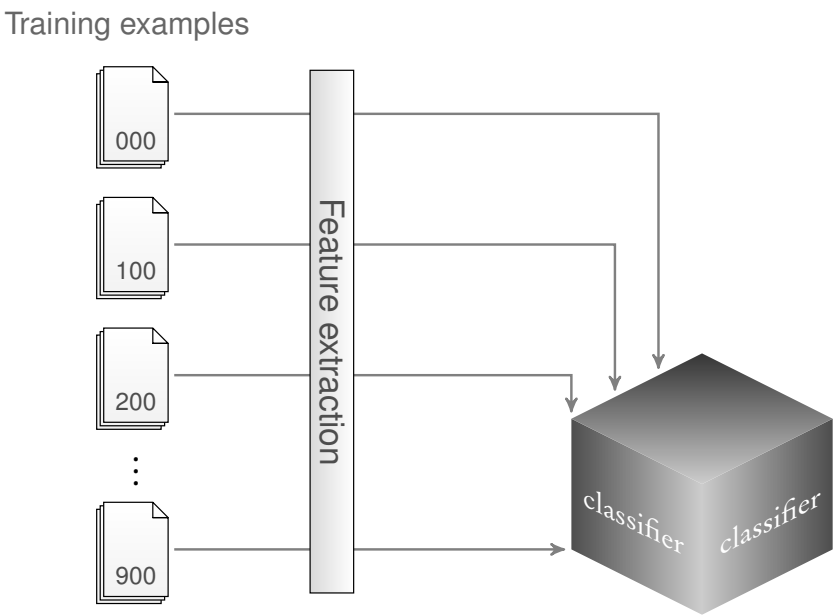
Ziel: Programm soll aus dem Text eines Dokuments die richtige Bedeutung (= Zielklasse) ableiten.

Methode: Maschinelles Lernen

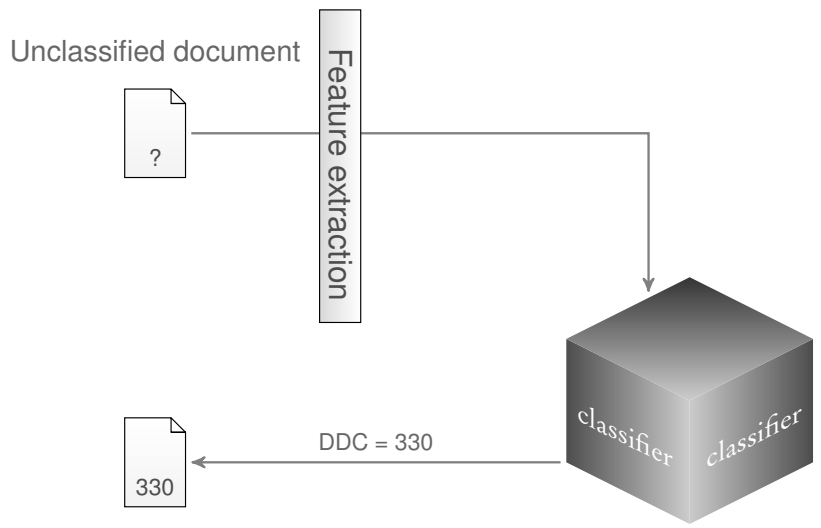
- Automatische Generierung eines Klassifikators aus Beispieldokumenten
- Lernen von Konzepten durch extensionale Beschreibung (= Aufzählung von Beispielen)

Textklassifikation durch einen überwachten Lerner

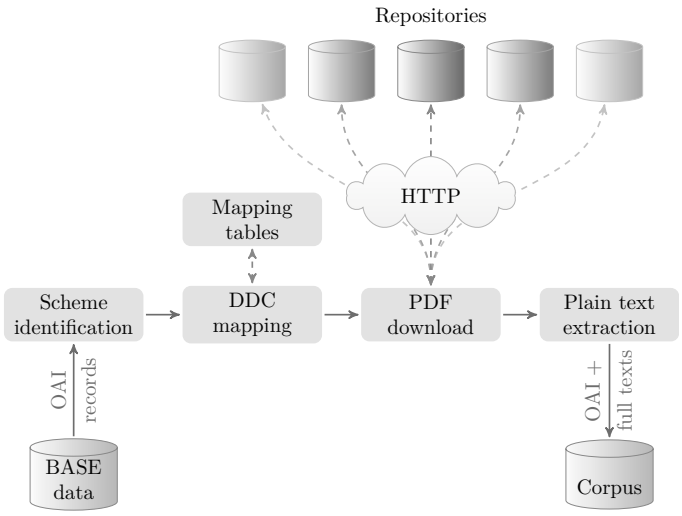
Lernphase



Applikationsphase

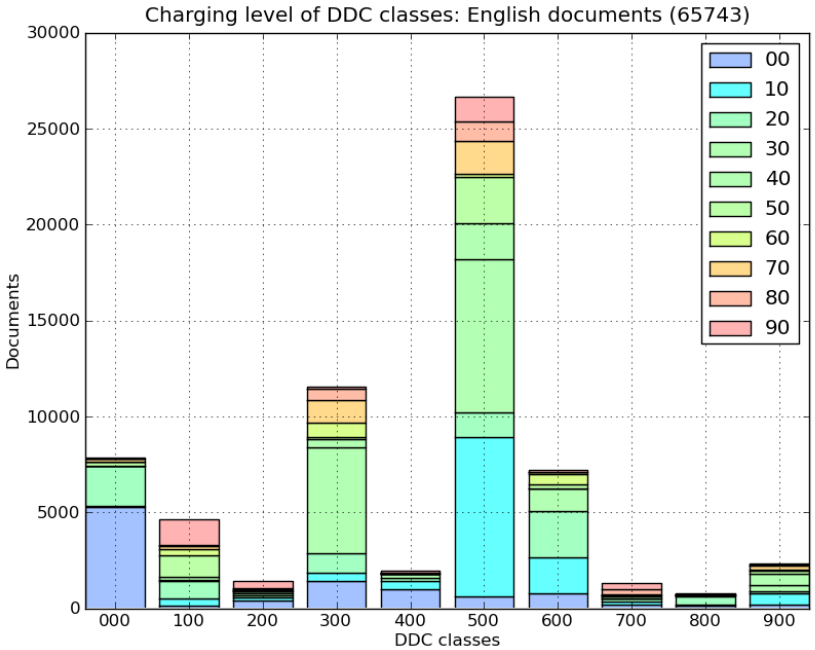


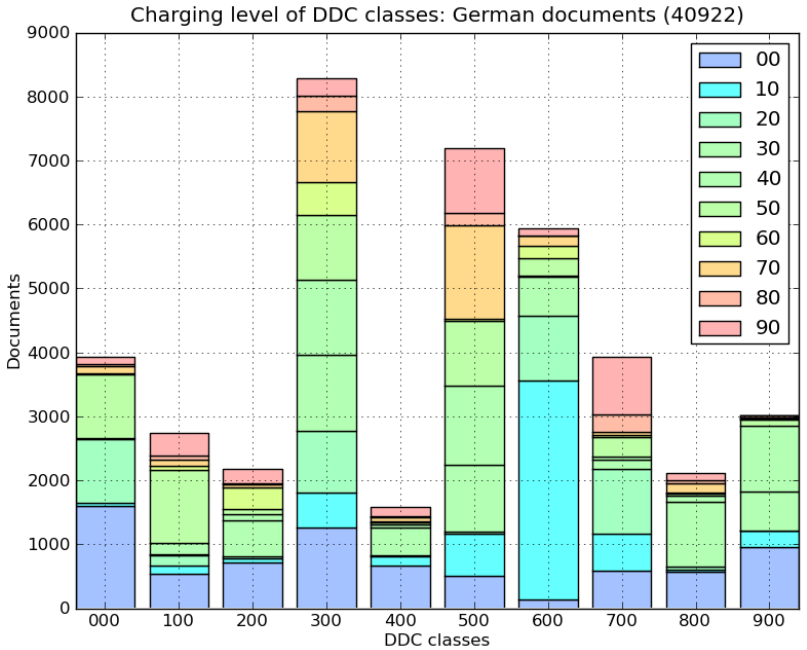
Trainingsdokumente



- Konstruktion eines DDC-kategorisierten Textkorpus aus der BASE-Datenbasis
- Metadaten + Volltexte
- ~ 100.000 Dokumente
- Deutsch und Englisch
- semi-automatische Vergabe von DDC-Nummern durch Konkordanzen zu Fachklassifikationen

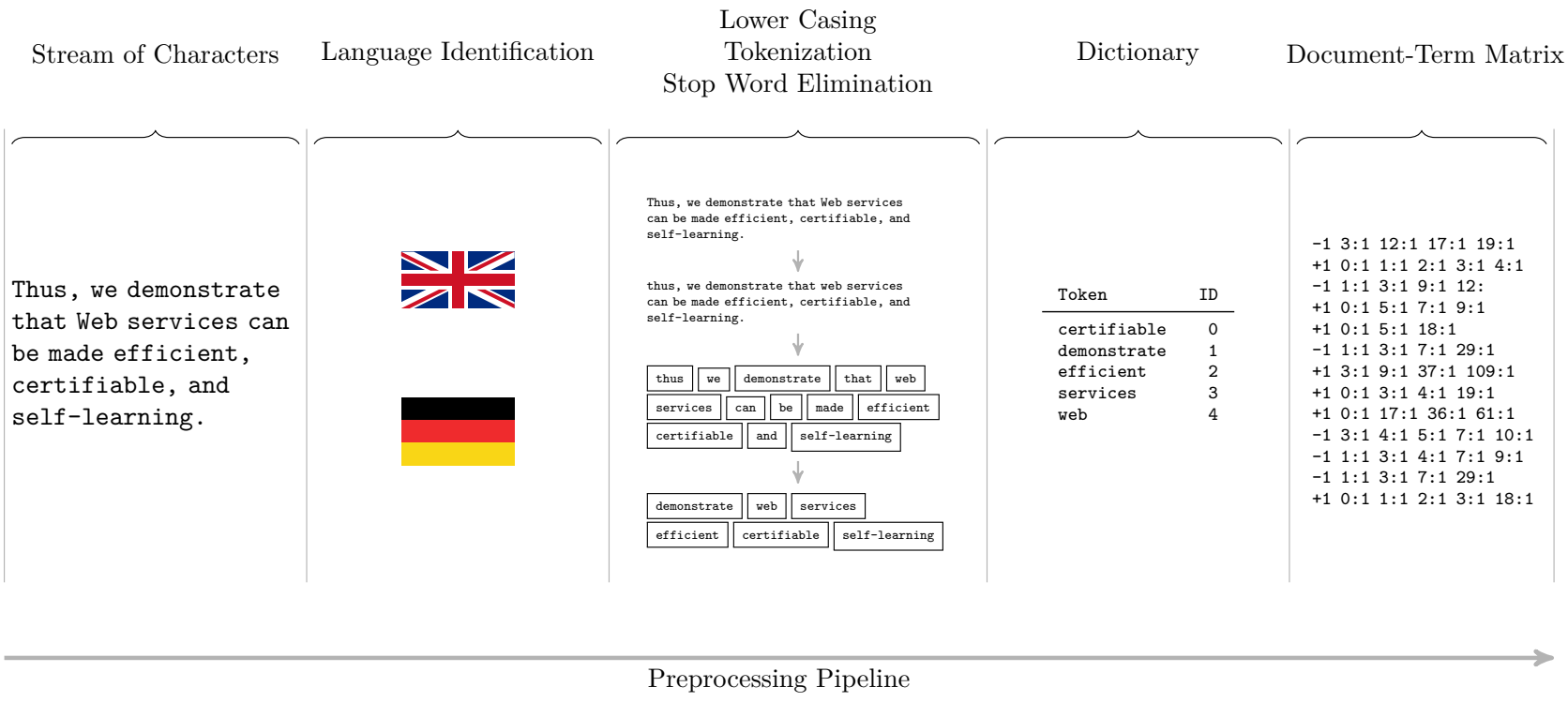
Probleme bei der Korpuserstellung





- Schiefe Verteilung der Dokumente über die DDC-Klassen
- Wenig Beispieldokumente in den Geisteswissenschaften
- Dokumentakquise ab der dritten DDC-Ebene (1.000 Klassen) extrem aufwändig mangels guter Sacherschließungsinformationen

Featureextraktion



Klassifikator

- Support Vector Machines (SVM)
- Vorteile
  - Gute Klassifikationsgenauigkeit
  - Schnell bei hochdimensionalen Problemen (z.B. Textklassifikation)
- Nachteile
  - SVM kann immer nur zwei Klassen lernen (bei mehr als zwei Klassen müssen zusätzliche SVMs trainiert werden)
  - Offline-Lerner: Neu-Training nötig bei Aktualisierung der Wissensbasis

Aktuelle Ergebnisse

Demo

- Klassifikationsgenauigkeit auf den ersten beiden DDC-Ebenen bis zu 90%
- testweise Anreicherung von bisher nicht-klassifizierten Dokumenten mit DDC-Nummern in BASE (derzeit ca. 50.000)

1 Motivation

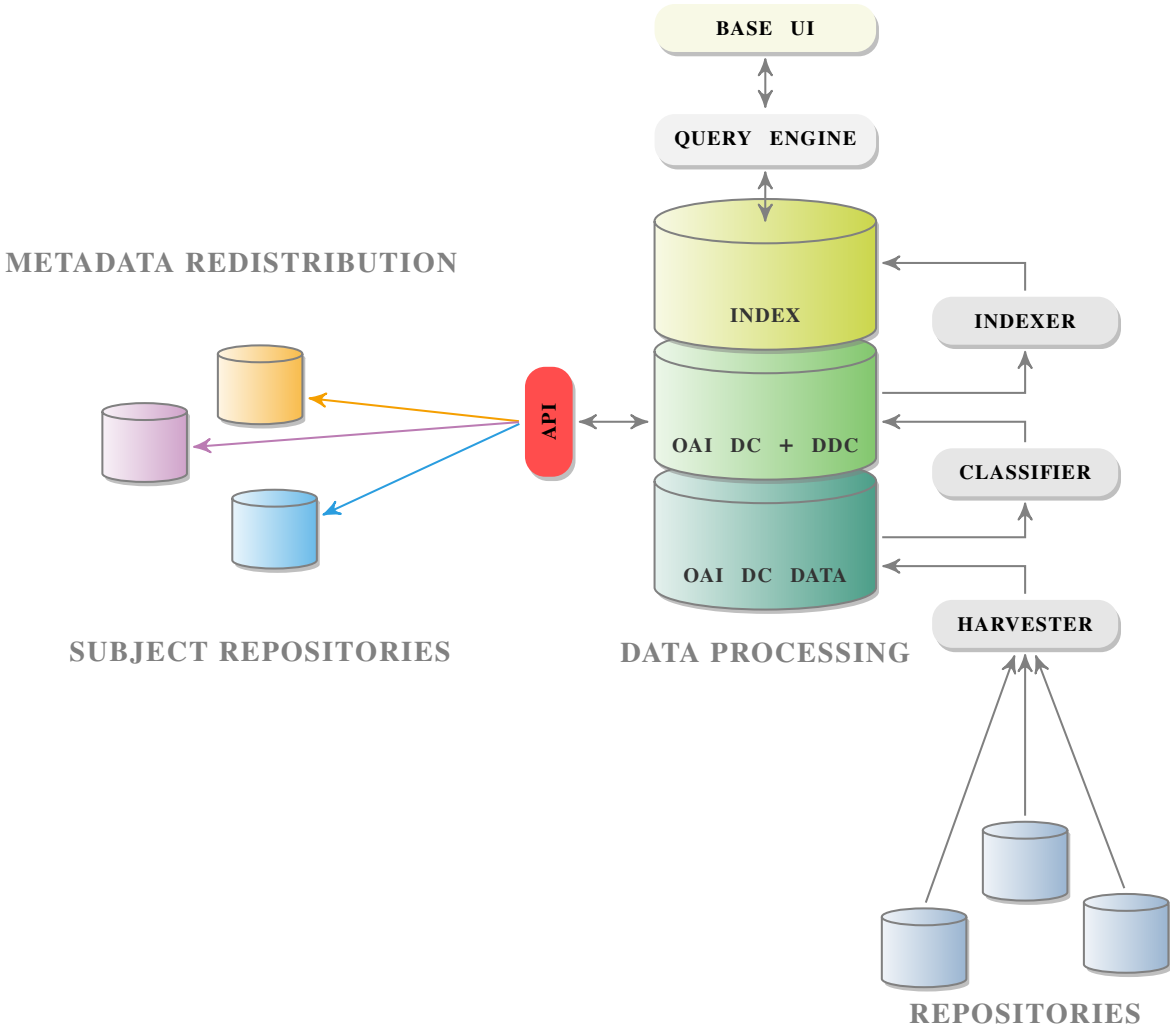
2 Automatische Klassifikation

3 **Use Cases**

4 Zusammenfassung

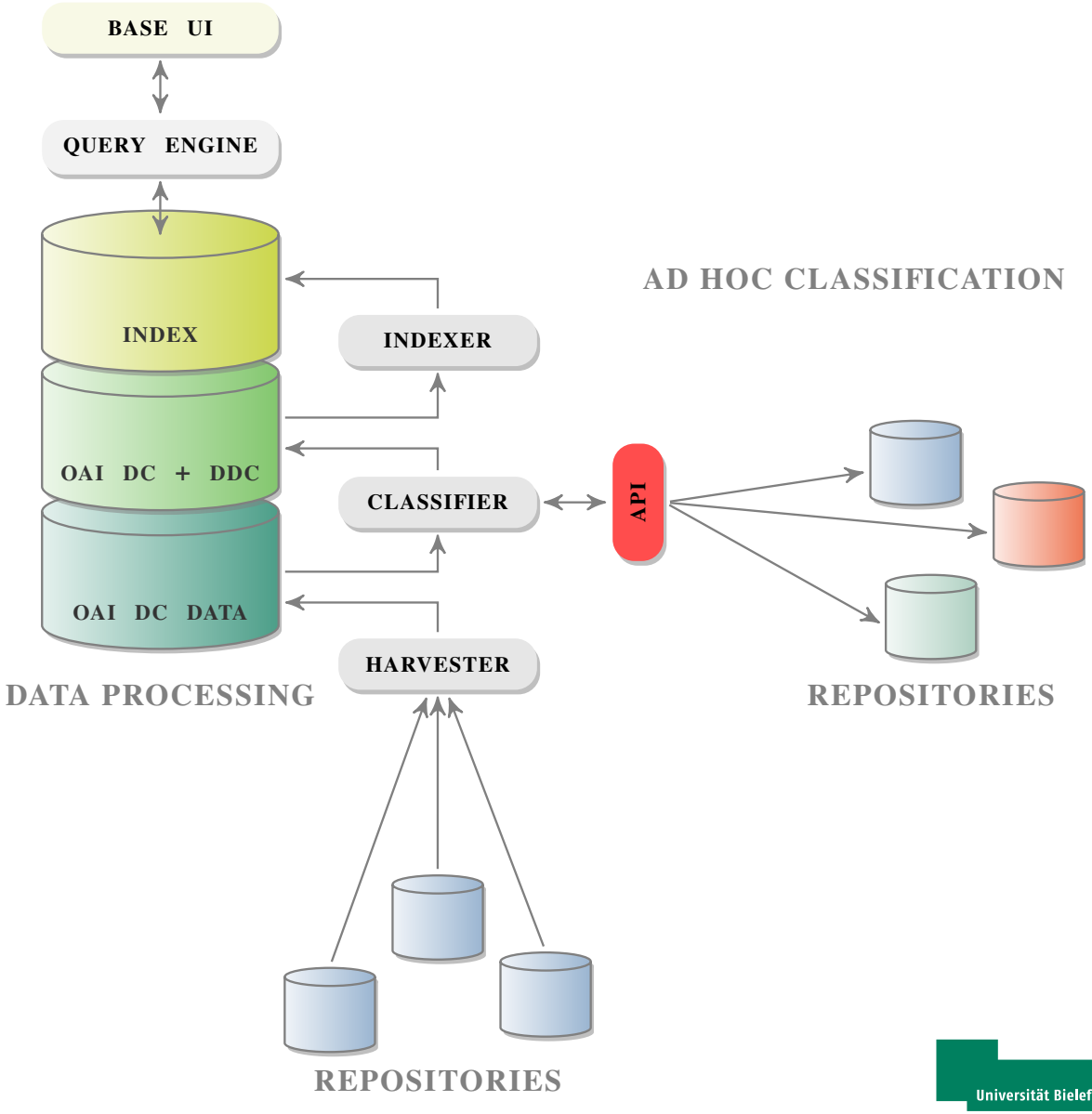
Use Cases

Belieferung von Fachrepositorien



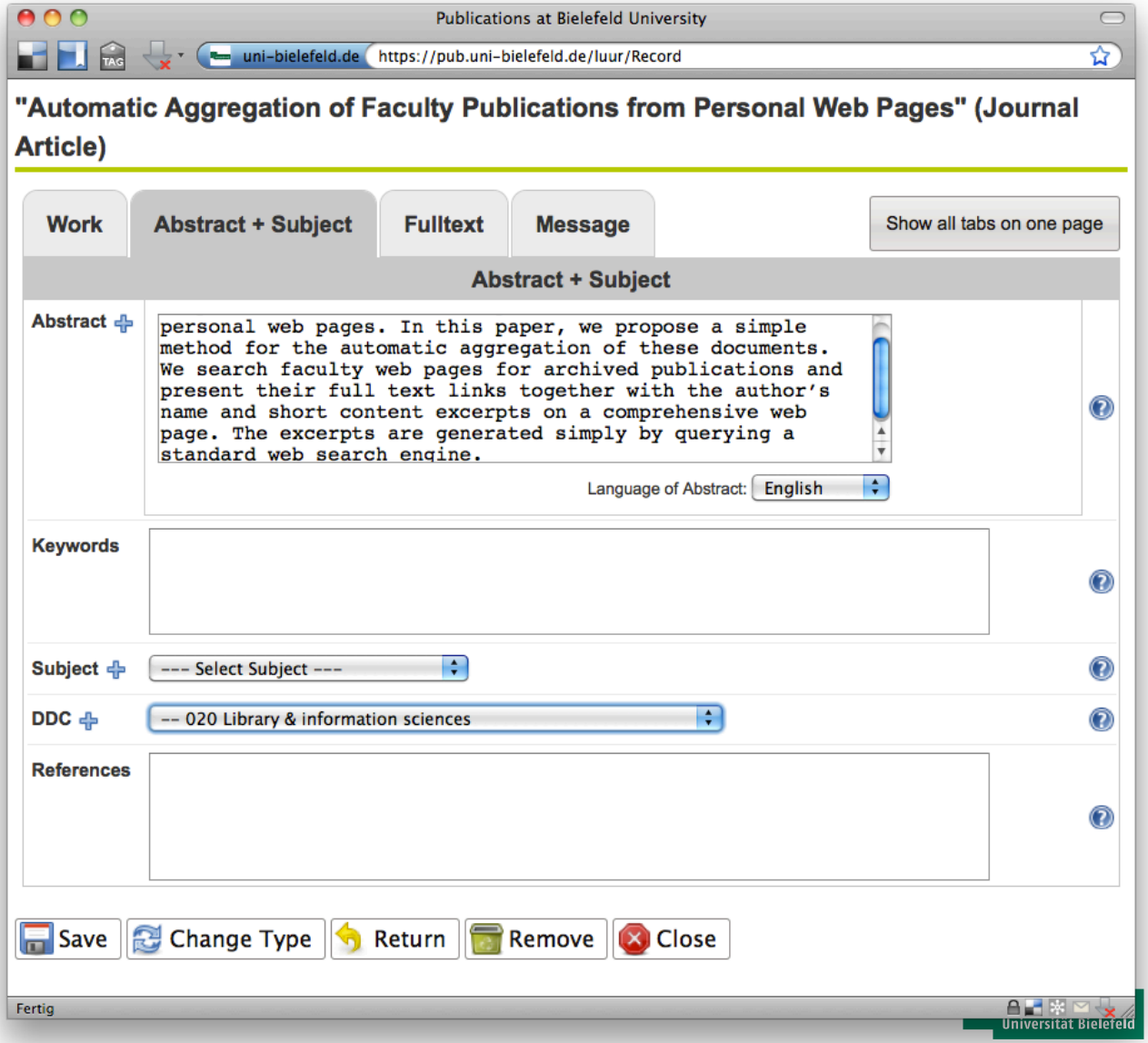
Use Cases

Verbesserung der Sacherschließung in institutionellen Repositorien



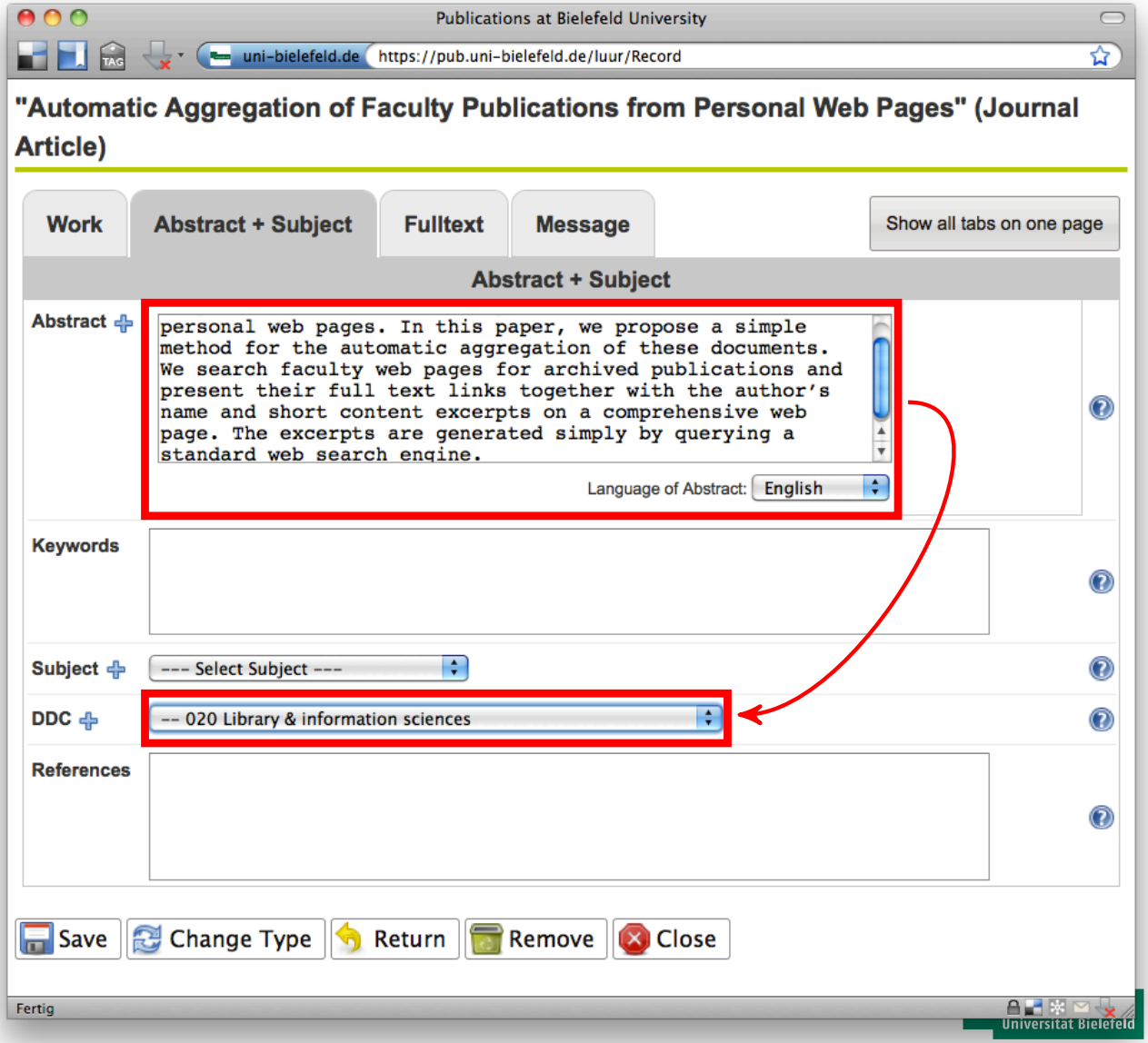
Use Cases

Vorschlagsystem für die Metadatenerfassung



Use Cases

Vorschlagsystem für die Metadatenerfassung



1 Motivation

2 Automatische Klassifikation

3 Use Cases

4 **Zusammenfassung**

Vorläufige Projekterkenntnisse

- Schwierigkeiten
  - Akquise von Trainingsdaten
  - Ab DDC Ebene 3: Abdeckung problematisch
- Erfolge
  - Grobklassifikation (1. und 2. Ebene) gut automatisierbar
  - automatische Vergabe von DNB-Sachgruppen (DINI-Empfehlung) auf jeden Fall erreichbar
  - semi-automatische Verfahren (Vorschlagssysteme) realistisch und sinnvoll
- Ausblick
  - Verbesserung des Klassifikators: Bündelung mehrerer Klassifikatoren (Boosting), Verbesserte Termauswahl (Terminologieextraktion)
  - Erforschung neuer Zielklassifikationen (z.B. Wikipedia-Kategoriensystem)

Vielen Dank für die Aufmerksamkeit!

Mathias.Loesch@uni-bielefeld.de