

Automatische Klassifikation elektronischer Dokumente

Mathias Lösch

Universitätsbibliothek Bielefeld

`Mathias.Loesch@uni-bielefeld.de`

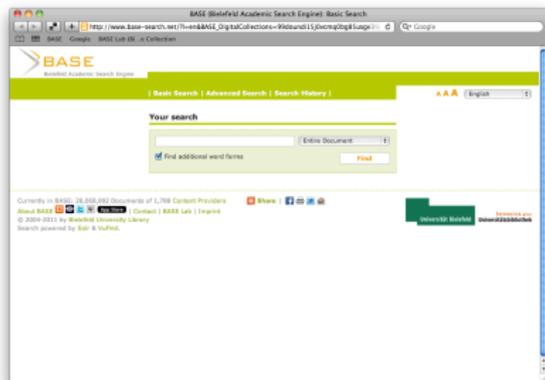
Kolloquium Wissensinfrastruktur, 20. Mai 2011

- DFG-Projekt »Automatische Anreicherung von OAI-Metadaten«
- Förderung Oktober 2009–September 2011
- Partner:
 - Universitätsbibliothek Bielefeld
 - Abteilung für geisteswissenschaftliche Fachinformatik, Universität 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



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

Motivation

BASE Browsing

BASE Lab (Bielefeld Academic Search Engine): Browse the Collection

http://lab.base-search.net/vufindtest/Browse/Dewey

BASE Google

BASE LAB
Bielefeld Academic Search Engine

Login

Basic Search | Advanced Search | Search History

English

Home > Browse >

Choose a Column to Begin Browsing:

Column	Category	View Records
DDC	0 Computer science, information & general works (1538)	View Records
	1 Philosophy & psychology (300)	View Records
	2 Religion (166)	View Records
	3 Social sciences (2920)	View Records
	4 Language (152)	View Records
	5 Natural sciences & mathematics (4234)	View Records
	6 Technology (2592)	View Records
	7 The arts; fine & decorative arts (261)	View Records
00 Computer science, information & general works (739)	01 Bibliography (2)	View Records
	02 Library & information sciences (67)	View Records
	05 General serial publications (181)	View Records
	07 News media, journalism & publishing (7)	View Records
020 Library & information sciences (67)	09 Manuscripts & rare books (542)	View Records

Currently in BASE Lab: 26,943,055 Documents of 1,725 Content Providers

About BASE | Contact | BASE Lab | Imprint

© 2004-2011 by Bielefeld University Library
Search powered by Solr & VuFind.

Share | Facebook | Twitter | YouTube

Universitätsbibliothek Bielefeld
INFORMATION plus

Universität Bielefeld

- 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

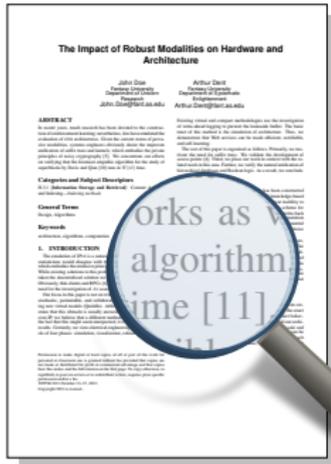
References

Save Change Type Return Remove Close

Fertig

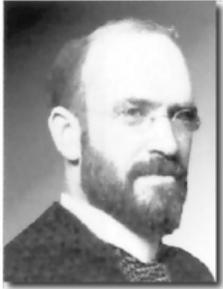
- 1 Motivation
- 2 Automatische Klassifikation**
- 3 Use Cases
- 4 Zusammenfassung

Automatische Klassifikation von Dokumenten auf Basis ihres Inhalts

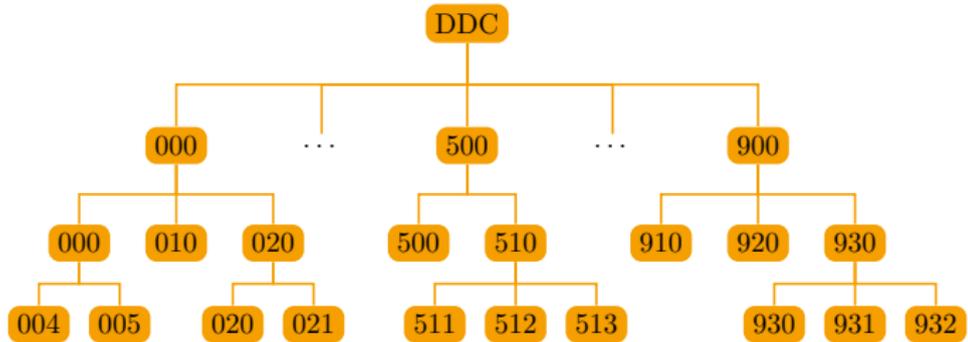


INFORMATIK

Dewey Decimal Classification



Melvil Dewey
(1851–1931)



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.

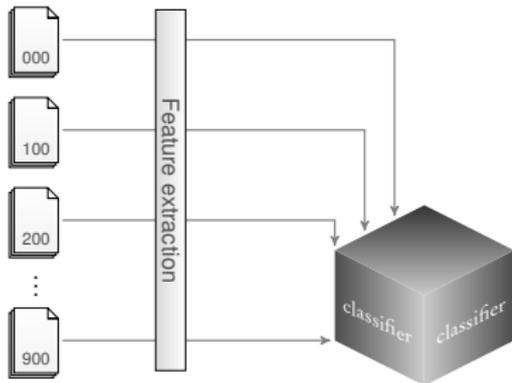
Methode: Maschinelles Lernen

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

Textklassifikation durch einen überwachten Lerner

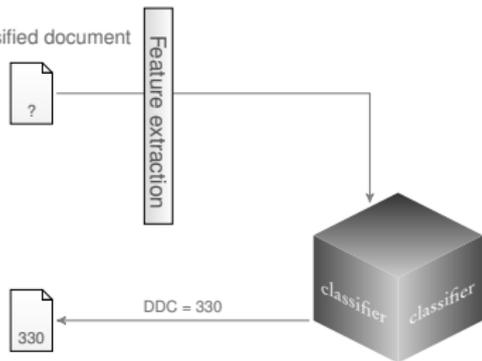
Lernphase

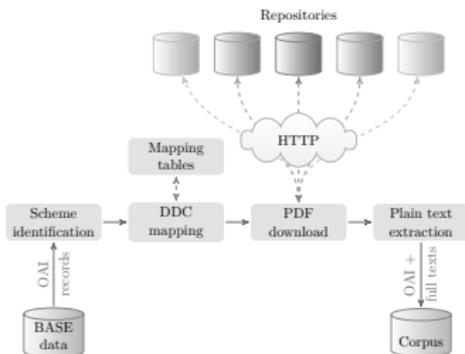
Training examples



Applikationsphase

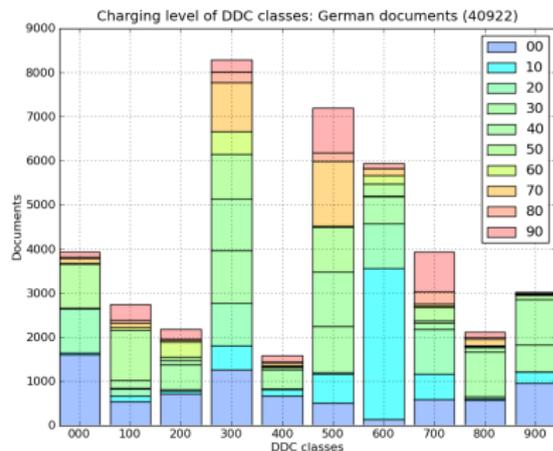
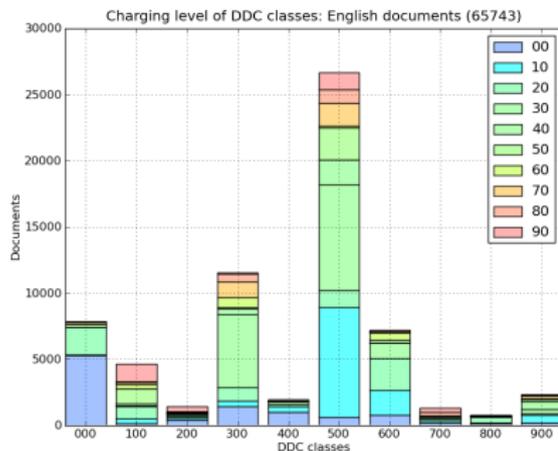
Unclassified document





- 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

Stream of Characters

Language Identification

Lower Casing
Tokenization
Stop Word Elimination

Dictionary

Document-Term Matrix

Thus, we demonstrate that Web services can be made efficient, certifiable, and self-learning.



Thus, we demonstrate that Web services can be made efficient, certifiable, and self-learning.

↓
thus, we demonstrate that web services can be made efficient, certifiable, and self-learning.

↓
thus we demonstrate that web services can be made efficient certifiable and self-learning

↓
demonstrate web services efficient certifiable self-learning

Token	ID
certifiable	0
demonstrate	1
efficient	2
services	3
web	4

-1 3:1 12:1 17:1 19:1
+1 0:1 1:1 2:1 3:1 4:1
-1 1:1 3:1 9:1 12:
+1 0:1 5:1 7:1 9:1
+1 0:1 5:1 18:1
-1 1:1 3:1 7:1 29:1
+1 3:1 9:1 37:1 109:1
+1 0:1 3:1 4:1 19:1
+1 0:1 17:1 36:1 61:1
-1 3:1 4:1 5:1 7:1 10:1
-1 1:1 3:1 4:1 7:1 9:1
-1 1:1 3:1 7:1 29:1
+1 0:1 1:1 2:1 3:1 18:1

Preprocessing Pipeline

- 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

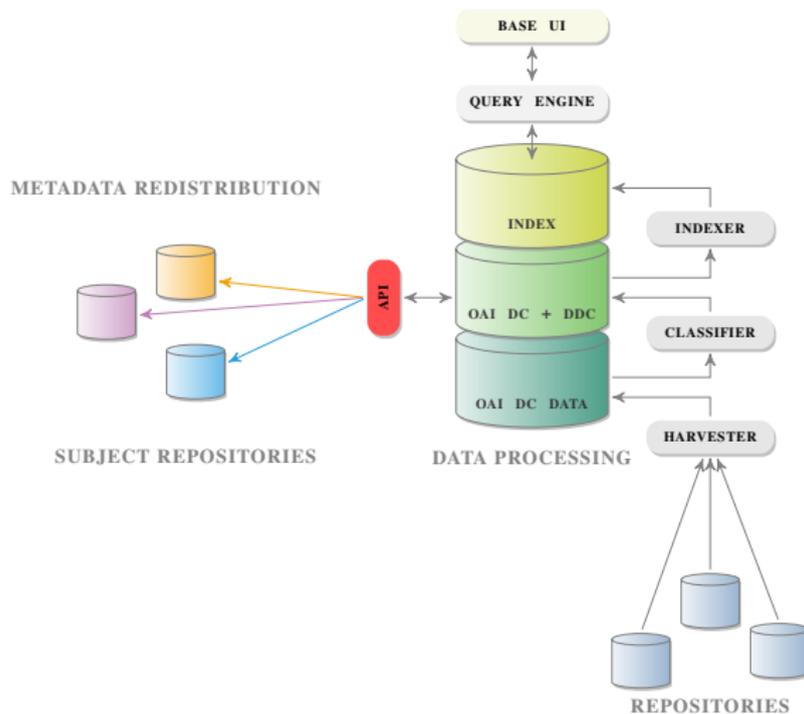
Demo

- 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

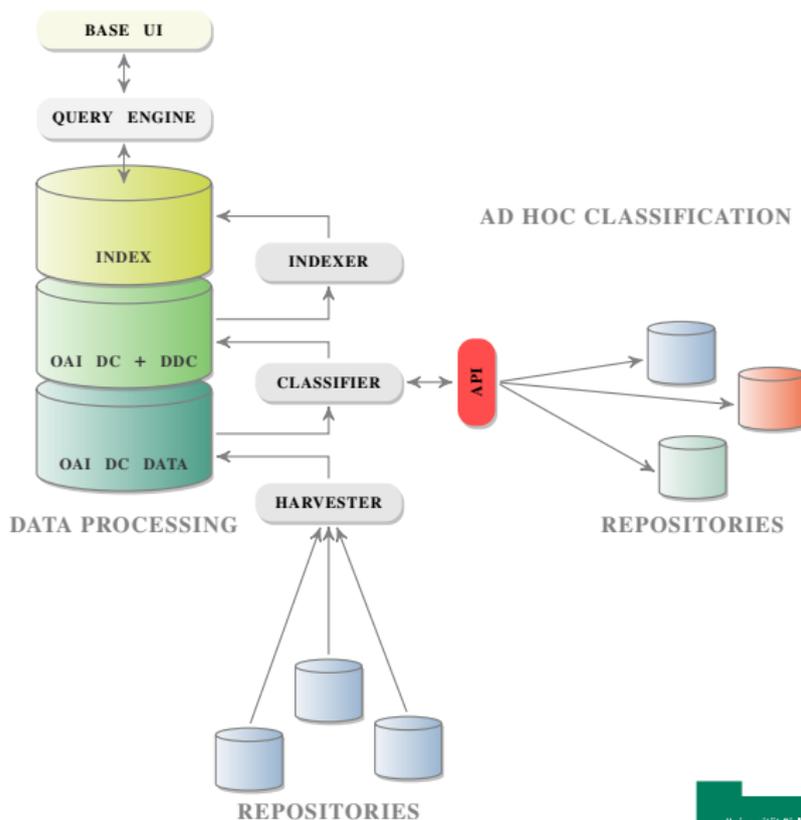
Use Cases

Belieferung von Fachrepositorien



Use Cases

Verbesserung der Sacherschließung in institutionellen Repositorien



Use Cases

Vorschlagsystem für die Metadatenerfassung

The screenshot shows a web browser window titled "Publications at Bielefeld University" with the URL "https://pub.uni-bielefeld.de/luur/Record". The main heading is "Automatic Aggregation of Faculty Publications from Personal Web Pages" (Journal Article). Below the heading are four tabs: "Work", "Abstract + Subject", "Fulltext", and "Message". A button "Show all tabs on one page" is located to the right of these tabs. The "Abstract + Subject" tab is active, displaying the following fields:

- Abstract:** A text area containing the text: "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." Below the text is a "Language of Abstract" dropdown menu set to "English".
- Keywords:** An empty text input field.
- Subject:** A dropdown menu with the text "--- Select Subject ---".
- DDC:** A dropdown menu with the text "-- 020 Library & information sciences".
- References:** An empty text input field.

At the bottom of the interface, there are five buttons: "Save", "Change Type", "Return", "Remove", and "Close". The status bar at the bottom left shows the word "Fertig". The bottom right corner features logos for "UNIVERSITÄT BIELEFELD", "INFORMATIONSSYSTEME", "Universitätsbibliothek", and "DFG".

Use Cases

Vorschlagsystem für die Metadatenerfassung

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

UNIVERSITÄT BIELEFELD

INFORMATIONSSYSTEME
Universitätsbibliothek

DFG

- 1 Motivation
- 2 Automatische Klassifikation
- 3 Use Cases
- 4 Zusammenfassung**

- 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