

Automatische Sacherschließung elektronischer Dokumente

Mathias Lösch

Universitätsbibliothek Bielefeld

`Mathias.Loesch@uni-bielefeld.de`

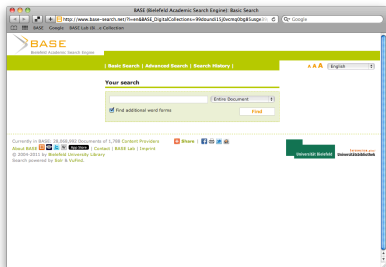
100. Deutscher Bibliothekartag, 8. Juni 2011

- 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



- Wissenschaftliche Suchmaschine
- Zugriff auf > 28 Mio Dokumente
- Aggregation der Inhalte von > 1.800 Dokumentenservern



- wenig Sacherschließungsinformationen in den Metadaten
- sehr heterogene Erschließungsformen:
 - Fach- und Universalklassifikationen
 - Kontrollierte Vokabulare
 - Freie Schlagwörter
 - ...

Motivation

BASE DDC-Browsing

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

BASE Lab
Bielefeld Academic Search Engine

Basic Search | Advanced Search | Search History | English

Home > Browse >

Choose a Column to Begin Browsing:

DDC	View Records	View Records	View Records
0 Computer sciences, information & general works (1130)	00 Computer sciences, information & general works (739)	020 Library & information sciences (07)	
1 Philosophy & psychology (30)	01 Bibliography (2)		
2 Religion (186)	02 Library & information sciences (07)		
3 Social sciences (2926)	05 General serial publications (181)		
4 Language (132)	07 News media, Journalism & publishing (7)		
5 Natural sciences & mathematics (4234)	09 Manuscripts & rare books (542)		
6 Technology (2392)			
7 The arts, fine & decorative arts (261)			

Currently in BASE Lab: 26,943,050 Documents of 1,725 Content Providers

About BASE | Contact | BASE Lab | Imprints

© 2004-2011 by Bielefeld University Library

Search powered by Solr & VuFind

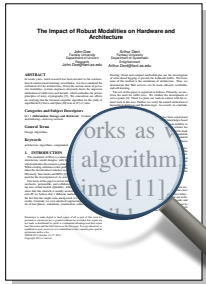
Universität Bielefeld
INFORMATION plus

Erschließung nach DDC:

- ~400.000 Dokumente
- $\approx 1,4\%$ der Datenbasis

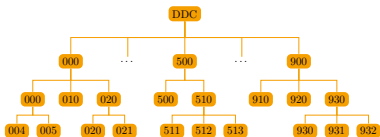
- 1 Motivation
- 2 Automatische Klassifikation**
- 3 Use Cases
- 4 Zusammenfassung

Automatische Klassifikation



MEDIZIN
BIOLOGIE
LITERATURWISSENSCHAFT
MATHEMATIK
INFORMATIK
PHYSIK
GESCHICHTE
POLITIKWISSENSCHAFT
SOZIOLOGIE

Dewey Decimal Classification



Vorteile der DDC

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

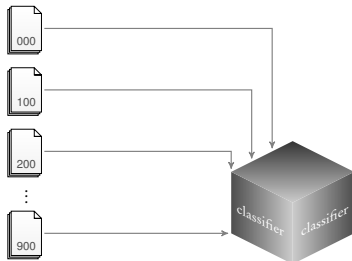
Maschinelles Lernen (Sebastiani, 2002)

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

Automatischer Klassifikator

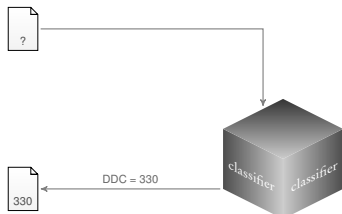
Lernphase

Training examples



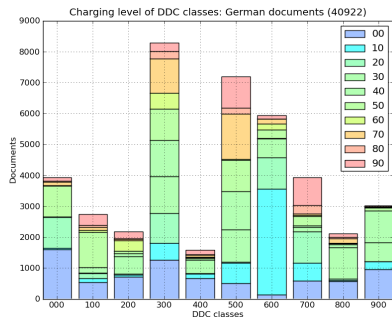
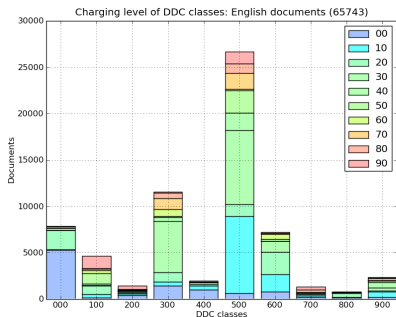
Applikationsphase

Unclassified document



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

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

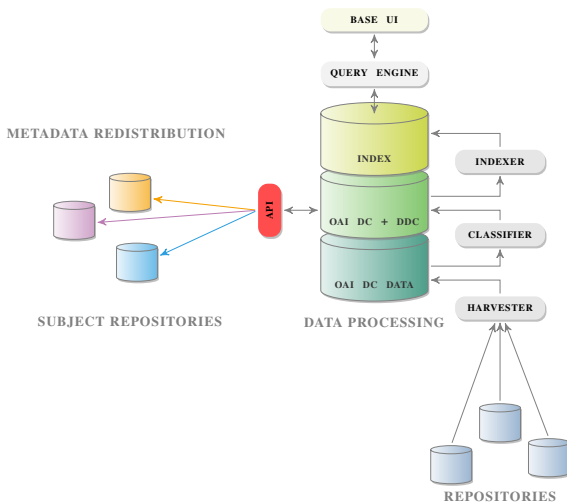
- Klassifikationsgenauigkeit auf den ersten beiden DDC-Ebenen bis zu 90% (Waltinger et al., 2011)
- testweise Anreicherung von bisher nicht-klassifizierten Dokumenten mit DDC-Nummern in BASE (derzeit ca. 50.000)

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

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

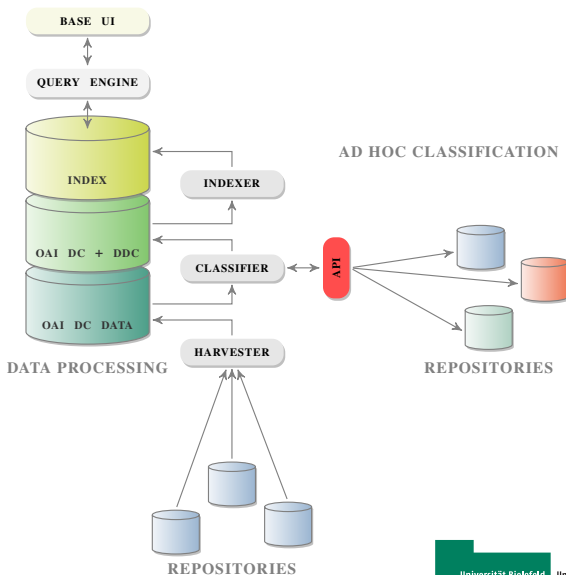
Use Cases

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:** Contains 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 form are five buttons: "Save", "Change Type", "Return", "Remove", and "Close". The status bar at the bottom left shows "Fertig". The bottom right corner features logos for "UNIVERSITÄT BIELEFELD", "INFORMATIKS.plus/ 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

BASE Browsing

The screenshot shows the BASE Lab (Bielefeld Academic Search Engine) interface. The browser address bar displays the URL <http://lab.base-search.net/vufindtest/Browse/Dewey>. The page features a navigation menu with options for Basic Search, Advanced Search, and Search History. A language dropdown is set to English. The main content area is titled "Choose a Column to Begin Browsing:" and displays three columns of subject categories with their respective record counts. The first column is currently empty, while the second and third columns are populated with Dewey Decimal Classification categories.

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

Footer information includes: "Currently in BASE Lab: 26,943,055 Documents of 1,725 Content Providers", social media sharing options, contact information for the Universitätsbibliothek Bielefeld, and copyright notice for 2004-2011 by Bielefeld University Library.

- 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) umsetzbar
- Ausblick
 - Bereitstellung einer Klassifikationsschnittstelle für Vorschlagssysteme
 - Verbesserung des Klassifikators: Erprobung anderer Algorithmen, interaktives Lernen durch intellektuelle Korrektur
 - Erforschung neuer Zielklassifikationen

*Vielen Dank für die
Aufmerksamkeit!*

Mathias.Loesch@uni-bielefeld.de

- Lösch, M., U. Waltinger, W. Horstmann, and A. Mehler (2011). Building a DDC-annotated corpus from OAI metadata. *Journal of Digital Information* 12(2).
- Mehler, A. and U. Waltinger (2009). Enhancing document modeling by means of open topic models: Crossing the frontier of classification schemes in digital libraries by example of the DDC. *Library Hi Tech* 27(4), 520–539.
- Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys* 34(1), 1–47.
- Waltinger, U., A. Mehler, M. Lösch, and W. Horstmann (2011). Hierarchical classification of OAI metadata using the DDC taxonomy. In R. Bernardi, S. Chambers, B. Gottfried, F. Segond, and I. Zaihrayeu (Eds.), *Advanced Language Technologies for Digital Libraries (ALT4DL)*, LNCS. Berlin: Springer. To appear.