



Automatische Vergabe von RVK-Notationen mittels fallbasiertem Schließen

Magnus Pfeffer
Universitätsbibliothek Mannheim



Gliederung

- Motivation und Historie
- Verfahren
- Umsetzung
- Experimentelle Ergebnisse
- Nachnutzung
- Ausblick



Motivation

- Einführung der RVK an der UB Mannheim
 - Unterstützung der Fachreferenten
 - Systematischer Zugang zum Gesamtbestand
- Wissen um Lernverfahren



Historie

- 2005: Erste Versuche
 - Einfaches Verfahren
 - Einspielung der Daten in den OPAC
- 2007: Systematische Untersuchung
 - Datenbasis UB Mannheim
 - Komplette Neuentwicklung
 - Masterarbeit (HU Berlin)
- 2008
 - Datenbasis Gesamtabzug Südwestverbund
 - Neue Verfahren



Verfahren: Fallbasiertes Schließen

■ Grundlagen

- Maschinelles Lernverfahren
- Prinzip: Ähnliches Problem - ähnliche Lösung

■ Ablauf

- Fall: Bekanntes Problem mit bekannter Lösung
- Speichern von Fällen in der Fallbasis
- Vergleich neues Problem - Probleme aus Fallbasis
- Ermittlung des ähnlichsten Problems und Lösung

■ Besonderheiten

- Keine inhaltliche Analyse
- Keine Ermittlung von Regeln oder Heuristiken



Verfahren

■ Vorteile

- Verfahren analog zur (Fremd)Datenübername
- Verfahren teilweise resistent gegen Widersprüche
- Fallbasis beliebig erweiterbar

■ Problembereiche

- Modellierung und Speicherung der Fälle
- Vergleichbarkeit der Probleme
- Effiziente Suche in der Fallbasis
- Komplexität steigt mit Größe der Fallbasis



Umsetzung

■ Modellierung

- Problem = Titelaufnahme
- Reduktion auf Titelwörter + Schlagwörter
- Lösung = Klassifikation

■ Speicherung

- Index Titelwörter
- Index Schlagwörter

■ Ähnlichkeit

- Grad der Übereinstimmung von Titelwörter + Schlagwörter
- Verfahren aus Stringvergleich
- Verfahren aus Information Retrieval



Umsetzung: Voraussetzungen

■ Inhaltliche Klassifikation

- Keine Zeitschriften
- Keine Reihen
- Keine Formalen Elemente

■ Datenqualität

- Gültige Klassifikationen
- Anzahl der Klassifikationen pro Titel



Umsetzung: Technisch

■ Daten

- Verbundabzüge in MAB2
- RVK-Struktur in XML

■ Aufbereitung

- Löschen von Zeitschriften
- Löschen von unselbständigen Einträgen
- Expansion der Reihen-GA in die Stücktitel
- Löschen von ungültigen Klassifikationen
- Löschen von formalen Klassifikationen



Ergebnisse

■ Testverfahren

- Neuklassifikation bereits klassifizierter Titel
- 1000 zufällige Titel
- Entfernen der Titel aus der Fallbasis

■ Bewertung

- Vergleich der automatischen und manuellen Klassifikation
- Gemeinsamer Vaterknoten im RVK-Baum
- Perfekt: Übereinstimmung
- Gut; Abstand 1-3
- Mäßig: Abstand >3 , aber noch gleiches Fach
- Schlecht: anderes Fach



Ergebnisse: Verfahren

■ Maximum

- Obere Schranke
- Alle Klassifikationen aller Titel mit Übereinstimmung(en)

■ Verfahren "Hamming"

- Basis: Stringvergleich
- $1 - [\#((A \cup B) - (A \cap B)) / \#A + \#B]$

■ Verfahren "IDF"

- Basis: Information Retrieval
- Summe der IDF aller übereinstimmenden terme



Ergebnisse

- Datenbasis UB Mannheim: 663.705 Titel

	theoretisches Maximum	Hamming	IDF
Perfekt	96,90%	48,80%	50,10%
Gut	2,60%	22,50%	22,30%
Mäßig	0,50%	9,60%	9,40%
Schlecht	0,00%	19,10%	18,20%
Median Klassifikationen	71497	4	4



Ergebnisse

- Datenbasis SWB: 2.496.839 Titel

	theoretisches Maximum	Hamming	IDF
Perfekt	98,30%	53,00%	54,80%
Gut	1,50%	20,40%	18,90%
Mäßig	0,20%	7,90%	7,70%
Schlecht	0,00%	18,70%	18,60%
Median Klassifikationen	134726.5	4	4

Ergebnisse nach Fachgebiet

■ Datenbasis SWB: 2.496.839 Titel

	A	B	C	D	E	F	G	H	I	K	L	M
Perfekt	56,9	64,4	63,6	59,1	52,7	71,5	62,7	63,7	60,0	61,2	56,1	52,5
Gut	18,4	15,3	20,1	21,0	15,8	13,9	14,5	14,4	15,9	13,8	13,9	21,6
Mäßig	4,3	7,8	1,8	1,8	5,6	4,0	7,6	10,8	11,0	11,2	8,3	3,1
Schlecht	20,4	12,5	14,5	18,1	25,9	10,6	15,2	11,1	13,1	13,8	21,7	22,8

	N	P	Q	R	S	T	U	V	W	X	Y	Z
Perfekt	60,0	54,7	58,4	47,3	72,1	60,6	69,6	69,9	59,7	0,0	0,0	0,0
Gut	13,7	24,1	23,0	11,4	18,2	12,0	16,7	11,4	16,0	0,0	0,0	0,0
Mäßig	10,8	8,5	4,7	13,7	0,5	1,3	1,1	4,5	2,7	0,0	0,0	0,0
Schlecht	15,5	12,7	13,9	27,6	9,2	26,1	12,6	14,2	21,6	0,0	0,0	0,0



Nachnutzung

■ Einspielung Verbunddatenbank

- + Keine neue Software erforderlich
- + Übernahme durch Bibliotheken einfach
- Neue Felder in Verbunddatenbank
- Nur für Verbundteilnehmer
- Nur periodische Updates

■ Webservice

- + Immer aktuellste Daten und Verfahren
- + Unabhängig von Verbundteilnahme
- Neue Software erforderlich



Ausblick

■ Verfahren

- Umsetzung in C++
- Stemming und Wortzerlegung wieder aktivieren
- Prüfung von Vergleichsverfahren mit n-grams

■ Projekt

- Komplettlauf Verbundabzug
- Angebot Einspielung SWB
- Prototyp Webservice
- Prototyp Desktop-Software

■ <http://www.bib.uni-mannheim.de:8080/Classification>

Fragen

