Automatische Vergabe von RVK-Notationen
-
aktueller Stand

Magnus Pfeffer
Universitätsbibliothek Mannheim

Gliederung

- Motivation und Projektverlauf
- Verfahren und Umsetzung
- Analysen und Ergebnisse
- Nachnutzung der Daten
- Ausblick

Motivation

- Einführung der RVK an der UB Mannheim
  - Unterstützung der Fachreferenten
  - Systematischer Zugang zum Gesamtbestand
  - Abschätzung Platzbedarf
- Wissen um KI-Verfahren
  - Reiz der praktischen Anwendung

Projektverlauf

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

Projektverlauf

- 2008 Q2: Weiterentwicklung
  - Neue Verfahren
  - Datenbasis Gesamtabzug Südwestverbund
- 2008 Q3: Grid-Portierung
  - Übertragung auf BW-Grid Cluster Mannheim
  - 144 Knoten mit je 8 CPUs
  - Dauer eines Laufs ca. 10 Stunden

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 dessen Lösung
- Besonderheiten
  - Keine inhaltliche Analyse
  - Keine Ermittlung von Regeln oder Heuristiken

Verfahren

- Vorteile
  - Verfahren analog zur (Fremd)Datenübername
  - Verfahren teilweise resistent gegen Widerspüche
  - 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
  - Flache Liste mit Zuordnung PPN-RVKs
- Ähnlichkeit
  - Grad der Übereinstimmung von Titelwörter + Schlagwörter
  - Unterschiedliche Verfahren möglich

Umsetzung: Voraussetzungen

- Inhaltliche Klassifikation
  - Keine Zeitschriften
  - Keine Reihen
  - Keine formalen Klassifikationen
- 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
  - Auswahl 10000 zufälliger Titel
  - Entfernen dieser Titel aus der Fallbasis
- Bewertung
  - Vergleich der automatischen und manuellen Klassifikation
  - Suche des nächsten gemeinsamen 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 ∪ B) - (A ∩ B)) / #A + #B]
- Verfahren “IDF”
  - Basis: Information Retrieval
  - Summe der IDF aller übereinstimmenden Terme

Ergebnisse

- Datenbasis SWB: 2.496.839 Titel

	theoretisches Maximum	Hamming	IDF
Perfekt	98,30%	57,26%	56,89%
Gut	1,50%	18,99%	18,84%
Mäßig	0,20%	6,93%	7,51%
Schlecht	0,00%	16,82%	16,76%
Median Klassifikationen	134726	4	4

Ergebnisse nach Fachgebiet

- Datenbasis SWB, Testmenge 1000 Titel, IDF

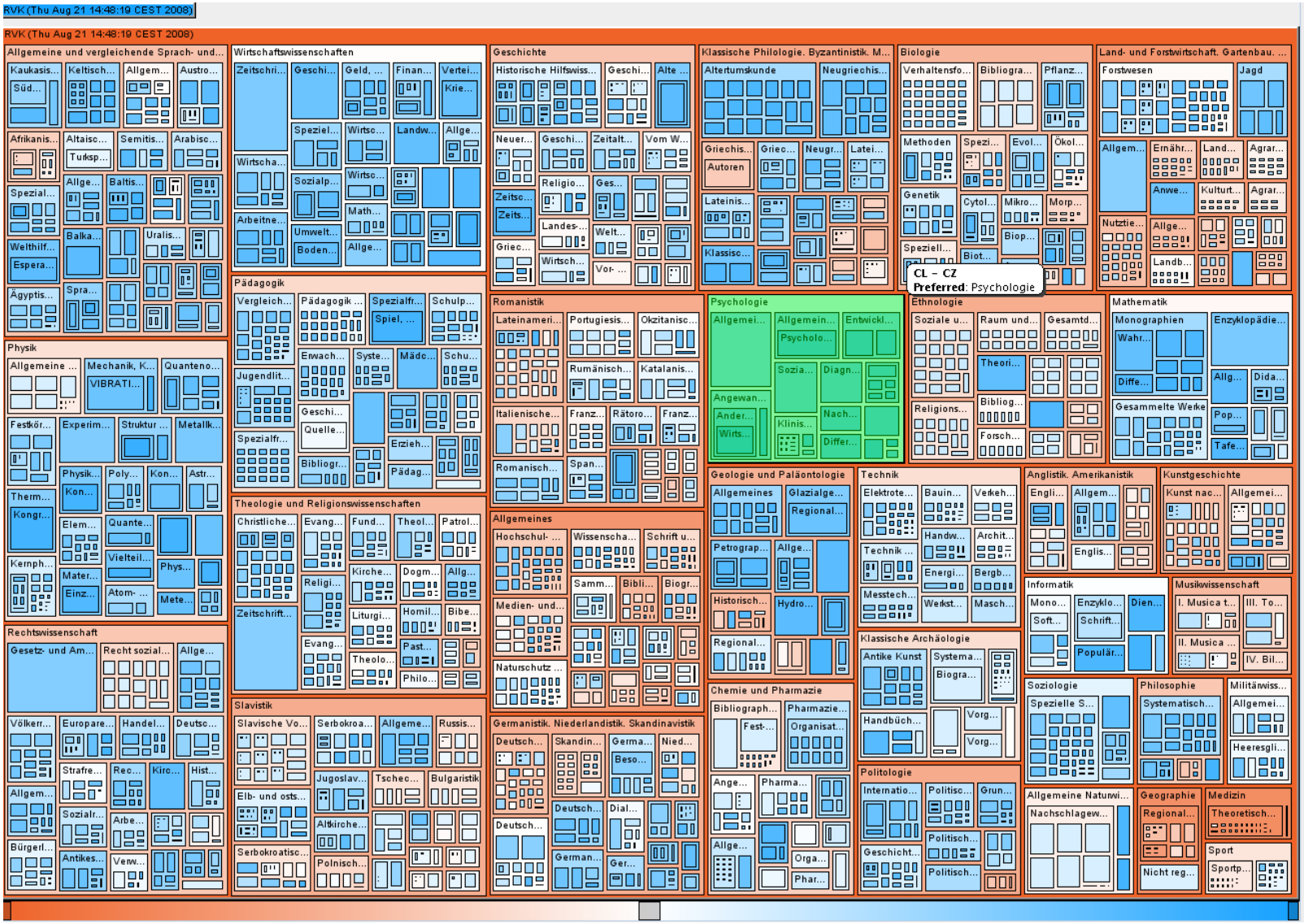
	A	B	C	D	E	F	G	H	I	K	L
Perfekt	56,9	64,4	63,6	59,1	52,7	71,5	62,7	63,7	60,0	61,2	56,1
Gut	18,4	15,3	20,1	21,0	15,8	13,9	14,5	14,4	15,9	13,8	13,9
Mäßig	4,3	7,8	1,8	1,8	5,6	4,0	7,6	10,8	11,0	11,2	8,3
Schlecht	20,4	12,5	14,5	18,1	25,9	10,6	15,2	11,1	13,1	13,8	21,7

	N	P	Q	R	S	T	U	V	W	X	Y
Perfekt	60,0	54,7	58,4	47,3	72,1	60,6	69,6	69,9	59,7	49,9	59,1
Gut	13,7	24,1	23,0	11,4	18,2	12,0	16,7	11,4	16,0	17,3	17,8
Mäßig	10,8	8,5	4,7	13,7	0,5	1,3	1,1	4,5	2,7	7,7	6,4
Schlecht	15,5	12,7	13,9	27,6	9,2	26,1	12,6	14,2	21,6	25,1	16,7

Tiefergehende Analyse

- Semtinel
  - Interaktive Analyse
  - Vollvergleich manuell zu automatisch
  - Ermittlung von „hot spots“ mit Fehlerhäufung
- Statistische Analysen
  - Korrelation von RVK-Klassen
  - Korrelation von Termen mit Klassen

Semtinel Treemap Visualisierung



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++ nicht mehr erforderlich
  - Prüfung von Vergleichsverfahren mit n-grams
  - Erwartungswert der Notationen deutlich reduzieren
    - Gestuftes Verfahren
    - Unterschiedliche Verfahren kombinieren
    - Einbeziehung weiterer Kategorien
    - Abbruch bei schlechter Datenlage
- Projekt
  - Einspielung SWB
  - Einbeziehung anderer Verbundkataloge

Ausblick

- Projekt
  - Einspielung in den Katalog des SWB
  - Einbeziehung anderer Verbundkataloge
  - „Rich Annotations“ mittels RDF für DC-Metadaten

Vielen Dank für Ihre Aufmerksamkeit

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