Automatische Vergabe von RVK-Notationen
Erfahrungen eines Projekts an der UB Mannheim

Dipl. Inform. Magnus Pfeffer, M.A.
magnus.pfeffer@bib.uni-mannheim.de

Universitätsbibliothek Mannheim
Automatische Vergabe von RVK-Notationen

Überblick

- Anlass des Projekts
- Grundlagen fallbasiertes Schließen
- Umsetzung und Implementierung
- Experimente
- Ergebnisse
- Mögliche Erweiterungen

Anlass des Projekts

- Größere Bibliotheksbereiche
- RVK zur gemeinsamen Aufstellung
- Unterstützung der Retrosystematisierung
- Unterstützung der Platzbedarfsplanung
- Virtuelle systematische Aufstellung im Katalog

Universitätsbibliothek Mannheim
Automatische Vergabe von RVK-Notationen

Fallbasiertes Schließen

- Maschinelles Lernverfahren
- Prinzip: ähnliches Problem – ähnliche Lösung
- Algorithmus
  - Aufbau Fallbasis mit bekannten Lösungen
  - Vergleich neuer Fall mit allen Fällen der Basis
  - Finden des ähnlichsten Falls der Basis
  - Adaption oder Übernahme von dessen Lösung
- Ohne Adaption: fallbasierte Klassifikation

Umsetzung auf RVK-Vergabe

- Probleme/Fälle
  - Titelaufnahmen ohne RVK-Notation
- Lösungen
  - Klassifikation
- Fallbasis
  - Bereits klassifizierte Titel
- Vergleich
  - Ähnlichkeitsmaß

Universitätsbibliothek Mannheim
Automatische Vergabe von RVK-Notationen

Annahmen

- Nur korrekte Notationen im Verbund
- Inhaltliche Klassifikation
  - ? RVK-Klassen mit formalen Kriterien
    - Zeitschriften
    - Reihen
    - Jahr der Veröffentlichung als Notationsbestandteil
- Eindeutige Klassifikation

Ähnlichkeitsmaß

- „Krux“ des Verfahrens
- Realisierung: Ähnlichkeitsfunktion
- Formale Kriterien
  - Selbstvergleich maximale Ähnlichkeit
  - Symmetrisch
  - Normiert
- Inhaltliche Kriterien
  - Berücksichtigung aller relevanten Daten
  - Gewichtung der Attribute

Universitätsbibliothek Mannheim
Automatische Vergabe von RVK-Notationen

Ähnlichkeitsfunktion

- Nur inhaltstragende Kategorien
- Titelwörter
  - Mehrsprachig
  - Zusammengesetzte Wörter
  - Flektierte Wörter
- Schlagwörter
  - Kontrolliertes Vokabular

Ähnlichkeitsfunktion

- Normierung der Titelwörter
  - Englisch
    - Endungen abschneiden
  - Deutsch
    - Wortzerlegung
    - Grundformbestimmung
- Vergleich von Wortmengen
  - Mehrfachauftreten nicht berücksichtigt

Universitätsbibliothek Mannheim
Automatische Vergabe von RVK-Notationen

Umsetzung

- Datenquelle
  - Verbundabzug im MAB2-Format
  - Extraktion von Titel- und Schlagwörtern
- RVK
  - XML-Datenabzug ? Baumdarstellung
  - Entfernung der problematischen Notationen
- Zerlegung und Normierung aller Wörter
  - Tools: Morphy, Snowball
  - Titelwörter ? Lexeme
- Aufbau von Indices
  - Titelwörter
  - Schlagwörter-IDs
  - Lexeme

Umsetzung

- Vergleich
  - Suche aller Elemente im Index
    ? Liste potentiell ähnlicher Titel
  - Direkter Vergleich mittels Ähnlichkeitsfunktion
- Retrieval
  - Klassifikation(en)?
    - Klassifikation(en) des ähnlichsten Titel
    - Häufigste Klassifikation(en) der n ähnlichsten Titel
    - Alle Klassifikationen der n ähnlichsten Titel
  - Absoluter Wert der Ähnlichkeit

Universitätsbibliothek Mannheim
Automatische Vergabe von RVK-Notationen

Ähnlichkeitsfunktion

- Simple: #(A n B) / Max(#A, #B)
  - Nur Übereinstimmungen
  - Symmetrisch und Normiert
- Hamming: 1 – [#((A U B) - (A n B)) / #A + #B]
  - Auch Nicht-Übereinstimmungen
  - Symmetrisch und normiert
- Edit: 1 – [#((A U B) - (A n B)) / #A + #B]
  - Aber unterschiedliche Gewichtung der Nicht-Übereinstimmungen
  - Nicht symmetrisch
  - Normiert

Experimente

- Testläufe Juni 2007
  - Masterarbeit HU Berlin
  - Verschiedene Ähnlichkeitsfunktionen
  - Verschiedene Retrievals
- Testverfahren
  - Klassifikation von 1000 Titeln mit Notationen (Goldstandard)?
  - Vergleichswert: Distanz im RVK-Baum

Universitätsbibliothek Mannheim
Automatische Vergabe von RVK-Notationen

Ergebnisse

- Theoretisches Maximum
  - Notationen aller Titel mit einem übereinstimmenden Element
- Zahlen
  - 94,7% korrekt (mindestens eine Notation identisch)?
  - 4,5% gut (minimale Distanz der Notationen: 1-3)?
  - 0,7% befriedigend (Notationen im gleichen Fachgebiet)?
  - 0,1% falsch (Notationen in unterschiedlichen Fachgebieten)?
  - Durchschnittlich 14950 Notationen

Ergebnisse

- Sieger
  - Funktion: Hamming
  - Elemente: Lexeme mit Schlagwörtern kombiniert
  - Retrieval: Notation(en) der/des ähnlichsten Titels
- Zahlen
  - 51,4% korrekt (mindestens eine Notation identisch)?
  - 22,7% gut (minimale Distanz der Notationen: 1-3)?
  - 11,1% befriedigend (Notationen im gleichen Fachgebiet)?
  - 14,8% falsch (Notationen in unterschiedlichen Fachgebieten)?
  - Durchschnittlich 5 Notationen
- Retrieval mit Häufigkeiten nahezu identisch

Universitätsbibliothek Mannheim
Automatische Vergabe von RVK-Notationen

Praktische Umsetzung UB Mannheim

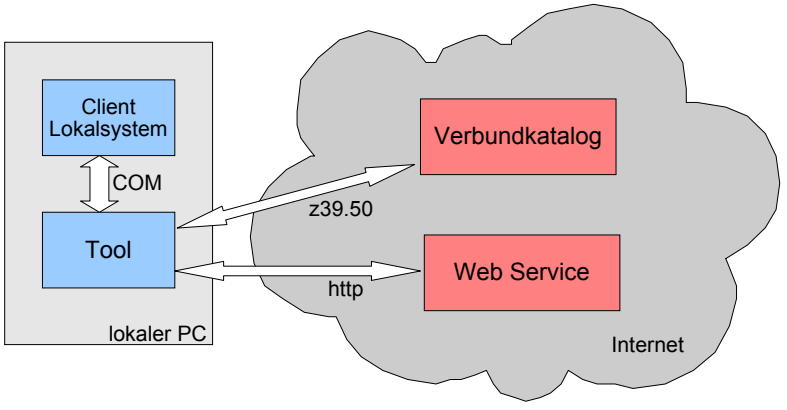
- Einspielung in Online-Katalog
  - Verbalisierung der Notation als Hilfe-Popup
  - Erstmals vollständiger systematischer Zugang
- Einsatz in der Retrosystematisierung
  - Nutzung durch Referenten
  - Titellisten nach RVK sortiert
  - Sehr hoher Nutzen
- Einsatz in der Bedarfplanung
  - Höhere Genauigkeit als reine Interpolation

Weitere Arbeiten

- RVK
  - Doppelklassen zusammenführen
  - Vollständiges Ausblenden der formalen Klassen
- Verfahren
  - Expansion der Schlagwörter-IDs zu Wörtern
  - Bessere Grundformzerlegung
  - Gewichtung der Terme nach Informationsgehalt
- Implementierung
  - schnellere Verarbeitung
  - Ziel: Annotations-Tool / Web Service

Universitätsbibliothek Mannheim
Automatische Vergabe von RVK-Notationen

Ausblick: Annotations-Tool



Fragen/Diskussion