

# **Fallbasierte automatische Klassifikation nach der RVK**

-

## **k-nearest neighbour auf bibliografischen Metadaten**

Magnus Pfeffer (Dipl.-Inform., M.A. LIS)  
Universität Mannheim, Universitätsbibliothek  
`magnus.pfeffer@bib.uni-mannheim.de`

# Themen

- Hintergrund und Motivation
- Ähnlichkeitsmaße
- Experimente und Ergebnisse
- User Feedback
- Aktuelle Entwicklungen

- Zusammenführung kleinerer Bereichsbibliotheken (2001)
  - Einführung der RVK als Aufstellungssystematik
  - Geringer Anteil an Fremddaten im Verbund
- Ziele
  - Beschleunigung des Erschließungsvorgangs durch Vorschlagssystem
  - Virtuelles Bücherregal im Katalog
    - Visualisierung der reklassifizierten Bestände
    - Grundlage für die Regalplanung
- Methode
  - Fallbasiertes Schließen
    - Übernahme der Klasse(n) der 1-nearest-neighbour
  - Datenbasis Südwestverbund

- Nur Metadaten

- Titel
- Personen und Körperschaften
- Schlagwörter

- Titelvergleich

- Zählen übereinstimmender Wörter

- Pro: Schnell, einfach
- Contra: gleiche Ähnlichkeitswerte bei identischen und erweiterten Titeln
  - z.B. “Einführung in Perl” maximal ähnlich zu “Einführung in den Compilerbau mit Perl”

- Alternative: Jaccard-Index

- Arbeitet auf Wortmengen
- Schnittmenge geteilt durch Vereinigungsmenge
  - Identität = 1, Nicht-ähnliche Titel = 0
  - jaccard (“Einführung in Perl”, “Einführung in den Compilerbau mit Perl”) =  $\frac{3}{5}$

- Gewichte
  - Stoppwortliste und Gleichgewichtung
    - Überraschend gute Ergebnisse
    - Lange Stoppwortliste verhindert falsch positive Ähnlichkeiten
  - Alternative: tf-idf
    - Ähnliche Ergebnisse
    - Stoppwortliste überflüssig
- Normalisierung
  - Stemming englischer Wörter
  - Teilwortzerlegung und Stammformreduzierung deutscher Wörter
  - Alternative: N-gramme

- Automatische Klassifikation bereits klassifizierter Titel
  - Titel wird für den Vorgang aus der Fallbasis ausgeblendet
  - Gewählte Variante: Jaccard mit Stoppwortliste und Gleichgewichtung, Stemming und Teilwortzerlegung
- Bewertung
  - Bestehende Klassifikation als “Goldstandard”
  - Ideal: Identische Notation wird gefunden
  - Noch gut: Nächste gefundene Notation ist nur wenige Knoten entfernt
- Ergebnisse (2008)
  - Je nach Fachgebiet zwischen 65 und 80% gute oder bessere Treffer
  - Aber:
    - Pro Titel werden recht viele Notationen geliefert
    - Keine Korellation der Güte mit den Ähnlichkeitsmaßen

- Anwender: Sacherschließer, Fachreferenten
- Nutzung: Retroklassifikation
- Erfahrungen
  - Zu viele “falsche” Notationen, Verwirrend
  - Starke Unterschiede in den Fächern
    - Informatik: Gut nutzbar
    - Jura: Nahezu nicht verwendbar
- Anpassungen
  - Reduktion der gelieferten Notationen
  - In der 1nn-Menge häufig auftretende Notationen werden präferiert
    - “Bester” Vorschlag oft nicht der häufigste
    - Keine wirkliche Verbesserung

- Höhere Anforderungen an Ähnlichkeit
  - Maß
    - Eine Übereinstimmung bei Autoren / Verfasser
    - Identischer Titel
    - Berücksichtigung des Einheitssachtitels (MAB Feld 304)
  - Ziel: unterschiedliche Ausgaben eines Werks zusammenführen
    - Unterschiedliche Auflagen und Drucke
    - Parallelausgaben in anderen Formaten
    - Nachdrucke (Verlagswechsel)
  - Umsetzung
    - Datenbasis Südwestverbund und HeBIS
    - Berücksichtigung von RVK und SWD-Schlagwörtern
    - Prüfung durch Sacherschließer der beteiligten Verbände

# Ausgaben zusammenführen

- Ausgangsdaten
  - SWB: 12,78 Mio. Berücksichtigte Titelaufnahmen, davon
    - 3,24 Mio. Titelaufnahmen mit RVK-Notation(en)
    - 3,98 Mio. Titelaufnahmen mit SWD-Schlagwörtern
  - Hebis: 8,84 Mio. Berücksichtigte Titelaufnahmen, davon
    - 1,93 Mio. Titelaufnahmen mit RVK-Notation(en)
    - 2,24 Mio. Titelaufnahmen mit SWD-Schlagwörtern
- Ergebnis (neuster Lauf 2011)
  - SWB
    - 959.419 Titel neu mit RVK
    - 636.462 Titel neu mit SWD
  - Hebis
    - 992.046 Titel neu mit RVK
    - 1.179.133 Titel neu mit SWD

## Real world example

- Freihandbestand Jura

- 55.445 Titel noch nicht reklassifiziert
- Ähnlichkeitsmaß wie beschrieben
- Datenquellen
  - SWB
  - Hebis
  - BVB (über z39.50)
- Ergebnis: 47649 mit RVK, 7796 ohne (86% Abdeckung)
- Zahlen für Mathematik und Geschichte ähnlich

- Aufbereitung für den Fachreferenten

- Bilden von kleinen Teilmengen mit inhaltlicher Kohärenz
  - Alte Systematik
  - Stichwörter
  - Schlagwörter
  - Autoren
- Lücken in den Vorschlägen werden durch Kontext schließbar

# Aufbereitung

File Edit View Navigate Tools Window Help



Links Arbeitsbereich x

- Amazon
- Google
- Verbund
- BVB (KVK)
- Katalog
- Primo



Irgendwo enthält fourier  Filterkonfiguration Filter

Autoren und Herausg...	Titel	Jahr	Auflage	Schlagwörter	Signatur	RVKs (mehrere Quellen)	Bearbeiterfeld	Letzte Änderung
Lanczos, Cornelius	Discourse on Fourier series	1966		Fourier-Reihe Harmonische Analyse	M Lanczos, C.: Discourse	SK 450	SK 450	28.02.2011 11:52:18
Sneddon, Ian Naismith	Fourier transforms	1951			M Sneddon, I. N.: Fourier	SK 450	SK 450	28.02.2011 11:53:40
Petersson, Hans	Konstruktion der Modulformen und der zu gewissen Grenzkreisgruppen gehoerigen automorphen Formen von	1950		Modulform Fourier-Koeffizient	M Petersson, H.: Konstruktion	AX 14200		
Kufner, Alois Kadlec, Jan	Fourier series	1971		Fourier-Reihe	M Kufner, A.: Fourier	SK 450	SK 450	28.02.2011 11:52:17
Calus, Irene M. Fairley, J. A.	Fourier series and partial differential equations A programmed course for students of science and	1970		Partielle Differentialgleichung Einführung	M Calus, I. M.: Fourier	SK 450 SK 540	SK 540	28.02.2011 11:50:06
Bracewell, Ronald N.	The Fourier transform and its applications	1965			M Bracewell, R.: The Fourier	SK 450	SK 450	28.02.2011 11:49:23
Lifermann, Jean	Théorie et applications de la transformation de fourier rapide	1977		Nachrichtentechnik Signalverarbeitung	M Lifermann, J.: Theorie		SK 450	28.02.2011 11:52:36
Goldberg, Richard R.	Fourier transforms	1970	Repr.	Fourier-Transformation Harmonische Analyse	M Goldberg, R.: Fourier	SK 450	SK 450	28.02.2011 11:51:39
Szmydt, Zofia	Fourier transformation and linear differential equations	1977	rev. and engl. transl.		M Szmydt, Z.: Fourier	SK 450 SK 520 SK 540	SK 450	28.02.2011 11:53:53
Föllinger, Otto	Laplace- und Fourier-Transformation	1977	1. Aufl.	Fourier-Transformation Laplace-Transformation	M Föllinger, O.: Laplace- und	SK 450	SK 450	28.02.2011 11:51:30
Filippi, Siegfried	Untersuchungen ueber die Fourier-Tschebyscheff-Approximation von	1970		Stammfunktion Fourier-	M Filippi, S.: Untersuchungen	SK 910 AV 20000	SK 910	28.02.2011 11:51:23
Ritt, Robert K.	Fourier series	1970			M Ritt, R. K.: Fourier	SK 450	SK 450	28.02.2011 11:53:35
Taibleson, Mitchell H.	Fourier analysis on local fields preliminary informal notes of university courses and seminars in mathematics	1975		Harmonische Analyse Lokaler Körper	M Taibleson, M. H.: Fourier	SI 670		
Beth, Thomas	Verfahren der schnellen Fourier-Transformation die allgemeine diskrete Fourier-Transformation - ihre	1984		Fourier-Transformation	M Beth, T.: Verfahren der	SK 450 SK 880	SK 450	28.02.2011 11:49:05
	roduction	1976		Zeitreihenanalyse Harmonische Analyse	M Bloomfield, P.: Fourier	QH 237 SK 450 SK 845 SK 845	SK 845	28.02.2011 11:49:16
	alysis an introduction	1969		Fourier-Transformation Harmonische Analyse	M Donoghue, W. F.: Distributions	SK 450	SK 450	28.02.2011 11:32:46
		1972		Harmonische Analyse Distribution	M Challifour, J. L.: Generalized	SK 600	SK 600	28.02.2011 11:50:09

Massenzuweisung RVK Farben Kontext

Signatur  M A    
 Jahr  1960

Picklist Window

analysis

# Weiteres Vorgehen

- Systematische Ausweitung des Verfahrens
  - Deutsche und internationale Quellen
  - Weitere Erschließungssysteme
    - DDC
    - LOCC
    - LOC-SH
    - ...
- Ausnutzung von Konkordanzen
  - Vorhandene aus diversen Projekten
  - Auswertung von Korrelationen
- Ergänzung der Lücken
  - 1-nearest neighbour, Jaccard, tf-idf, 4-gramme
  - Umsetzung in Java für Cluster (Hadoop / Mahout)

