



Äquivalenzklassen – Alle Doppelstellen der RVK finden

Magnus Pfeffer
Universitätsbibliothek Mannheim

`pfeffer@bib.uni-mannheim.de`



Überblick

- Einführung
- Motivation und Beispiel
- Analysen
- Ergebnisse
- Weiterführende Arbeiten
- Diskussion



Einführung

- Nicht zu übersehen sind auch Schwierigkeiten mit häufigen, vielleicht zu häufig möglichen Doppelstellen [...].

(Lorenz, Bernd (Hrsg.): Handbuch zur Regensburger Verbundklassifikation. Harrasowitz. 2003. Seite 35)



Motivation

- **Maschinelle Auswertung**
 - Automatische Klassifikation
 - Bestandsanalysen
 - Konkordanzerstellung

- **Benutzerschnittstelle der Online-Kataloge**
 - Virtuelles Bücherregal
 - Suchen ähnlicher Titel
 - Iteratives Einschränken von Ergebnismengen



Beispiel: Georg Wilhelm Friedrich Hegel (*1770, †1831)

- **BF 3900 - BF 3901** Hegel, Georg W. Fr.
 - Theologie / Philosophie
 - Kein Registereintrag
 - Unterteilung in 2 Klassen

- **CG 4060 - CG 4077** Hegel, Georg W.
 - Philosophie / Geschichte der Philosophie
 - Register: Hegel, Georg Wilhelm Friedrich
 - Unterteilung in 18 Klassen

- **DD 7040 - DD 7041** Hegel, Georg Wilhelm Friedrich
 - Pädagogik / Geschichte der Pädagogik und des Bildungswesens
 - Kein Registereintrag
 - Unterteilung in 2 Klassen



Beispiel: Georg Wilhelm Friedrich Hegel (*1770, †1831)

- **MC 6450 - MC 6453** Hegel, Georg Wilhelm Friedrich
 - Politologie / Geschichte der politischen Philosophie und der Ideologien
 - Register: Hegel, Georg W. / Politische Philosophie
 - Unterteilung in 4 Klassen
 - Bemerkung: s.a. BF 3900 f, CG 4060 ff., NB 4772

- **NB 4772** Hegel, Georg W.
 - Geschichte / Geschichte als Wissenschaft und Unterrichtsfach
 - Register: Hegel, Georg Wilhelm Friedrich
 - Keine Unterteilung

Beispiel: Georg Wilhelm Friedrich Hegel (*1770, †1831)

The screenshot shows the online catalog interface for the University of Mannheim library. The main content area displays the 'Vollanzeige des Titels' for 'Gesammelte Werke' by Hegel, published in 2008. A blue callout box points to the search criteria 'Regensbg. Klassifika: CG 4060' in the search results, with the text 'Suchanfrage: identische Notation'. Another blue callout box points to the 'CG 4060' notation in the 'RVK-Notationen' field, with the text 'Notation'. The interface includes a left sidebar with navigation options like 'Suchen', 'Konto', and 'Einstellungen'. The top navigation bar shows the library name and search services. The bottom of the page features a footer with the date '17.03.2009' and the text 'GfKI 2009, Dresden'.

PPN	287861074
Gesamttitel	Gesammelte Werke
Bandangabe	25,1
1. Autor	• Hegel, Georg Wilhelm Friedrich
2. Autor	• Bauer, Christoph Johannes [Hrsg.]
Titel	• Vorlesungen über die Philosophie des subjektiven Geistes
Zusatz	[Teilbd. 1], Nachschriften zu den Kollegien der Jahre 1822/23
Verfasserang.	Georg Wilhelm Friedrich Hegel. Hrsg. von Christoph Johannes Bauer
Ort	Hamburg
Verlag	• Meiner
Jahr	2008
Umfang	VI, 549 S.
ISBN	978-3-7873-1895-5
RVK-Notationen	• CG 4060
SFX: Volltextcheck / Services	UB MANNHEIM
Signatur / Bestellen	Alle Exemplare
Signatur / Bestellen	BB A3 i
Signatur / Bestellen	BB Schloss Ostflügel i
Übergeordn. Werk	Hegel, Georg Wilhelm Friedrich: Gesammelte Werke.



Inhaltliche Analysen

- **Auswertung der Klassenbeschreibung**
 - Abgrenzung: *siehe*
 - Ähnliche Klassen: *siehe auch*
 - Probleme:
 - Unvollständig, z.B. Hegel „siehe auch“ nur bei MC
 - Uneinheitliche Schreibweisen

- **Direkter Vergleich der Klassenbezeichnungen**
 - Suchen aller gleichlautenden Bezeichnungen
 - Probleme:
 - Uneinheitliche Bezeichnungen, z.B. bei Personennamen
 - Allgemeine Bezeichnungen, z.B. *Sekundärliteratur*
 - Schlüssel



Heuristische Analyse

- **Ansatz**
 - Basis: nach RVK erschlossene Titeldaten
 - Zählen des gemeinsamen Auftretens von RVK-Notationspaaren
 - Annahme: Korrelation der Anzahl mit der Wahrscheinlichkeit für inhaltlichen Bezug

- **Probleme**
 - Titel mit Inhalten aus mehreren Bereichen
 - Formale Klassifikation \Leftrightarrow Inhaltliche Klassifikation
 - Doppelstelle \Leftrightarrow Unscharfe Definition



Grundsätzliche Fragen

- Lokalisierung der Doppelstelle
 - Möglichst hoch in der Hierarchie
 - z.B. Philosophie / Autoren
 - Möglichst tief in der Hierarchie
 - z.B. Hegel, Georg W. / Primärliteratur
 - Gemeinsamer Nenner
 - z.B. Hegel, Georg W.

- Möglichkeit einer kombinierten Analyse
 - Inhaltliche Analyse korrekt, aber unvollständig
 - Vergleich der Bezeichnungen unvollständig und verzerrt
 - Heuristik vielleicht vollständig, aber stark verzerrt



Auswertung Vermerke

- Umsetzung
 - Basis: Komplettabzug der RVK im XML-Format (Version 2007)
 - Suchmuster:
 - s.a. [A-Z] [A-Z] [0-9] [0-9]
 - f und ff. gefiltert
 - Aufzählungen bis 10 Notationen
 - Besonderheiten des gerichteten Graphen (Baumstruktur)
 - Unterscheidung Blattknoten / innerer Knoten
 - Vervollständigung der Verweise zu Paaren
 - Experiment: Expansion der Verweise der Elternknoten in die Blattknoten
 - Paare von Notationen anstelle Paare von Knoten



Auswertung Vermerke

- Ergebnis
 - *siehe auch*
 - 236 Paare
 - Expandiert: 4885 Paare
 - *siehe*
 - 2388 Paare
 - Expandiert: 18545 Paare
- Hoher Anteil an inneren Knoten
- Doppelstellen lokalisiert beim gemeinsamen Nenner
- Expansion prinzipiell möglich



Auswertung Vergleich

- Umsetzung
 - Basis: Komplettabzug der RVK im XML-Format (Version 2007)
 - Invertierter Index Bezeichnung → Notationen
 - Verwerfen von Bezeichnungen mit mehr als 10 Notationen
 - Hohe Wahrscheinlichkeit für Schlüssel oder allgemeine Begriffe

- Ergebnis
 - 13.987 Bezeichnungen mit mehreren Notationen
 - 11.207 Bezeichnungen mit 2 bis 10 Notationen
 - Entspricht 58.844 Notationspaaren
 - Hoher Anteil an Blattknoten
 - Lokalisierung der Doppelstelle beim gemeinsamen Nenner, aber auch in mehreren Ebenen der Hierarchie



Auswertung Heuristik

- Umsetzung
 - Basis: SWB-Verbundabzug (Mitte 2008)
 - 2,5 Millionen Titel mit RVK-Notation(en)
 - 700.000 Titel mit mehreren Notationen
 - Summieren des Auftretens von Notationspaaren
 - n Titel mit RVK1, davon m auch mit RVK2
 - Schwellwerte für n und n/m
 - Ermittlung der Teilsystematik der beiden Notationen
 - Gleiche Teilsystematik → unscharfe Definition
 - Unterschiedliche Teilsystematik → Doppelstelle
 - Kombination der Notationstupel zu -paaren
 - Ein Tupel ausreichend



Auswertung Heuristik

- Ergebnis
 - Ohne Filterung
 - 586.061 Paare
 - Davon 299.838 mit Elementen in unterschiedlichen Fachsystematiken
 - Für Schwellwerte $n > 10$ und $n/m > 0.5$
 - 6411 Paare
 - Davon 3566 mit Elementen in unterschiedlichen Fachsystematiken

- Wahl des Schwellwertes willkürlich



Exemplarische kombinierte Analyse

- Menge A: Notationspaare der heuristischen Analyse
 - Sehr hohe Schwellwerte ($n > 20$; $n/m > 0.7$)
 - Elemente aus unterschiedlichen Teilsystematiken

- Menge B: Notationspaare der heuristischen Analyse
 - mittlere Schwellwerte ($n > 10$; $n/m > 0.4$)
 - Keine Einschränkung bezüglich der Elemente

- Menge C: Notationspaare aus dem Bezeichnungsvergleich
 - Verwerfen von Paaren mit innerem Knoten als Element

- Ergebnismenge = $A \cup (B \cap C)$



Kombinierte Analyse

- Ergebnismengen
 - A: 1275 Paare
 - B: 20197 Paare
 - C: 45394 Paare
 - $B \cap C$: 145 Paare
 - $A \cup (B \cap C)$: 1419 Paare



Vorläufiges Fazit

- Alle Verfahren liefern Paare mit inhaltlicher Beziehung
 - Intellektuelle Vorleistung in den Verweisen schwer zu extrahieren
 - Umsetzungen nur als Prototypen - Viele Verbesserungen denkbar
 - Heuristischer Ansatz vielversprechend

- Zusammenführung und Auswertung nicht trivial
 - Hauptproblem: unterschiedliche Lokalisierung der Doppelstellen

→ Bildung von Äquivalenzklassen möglich



Nächste Schritte

- Wiederholung mit aktueller RVK-Online XML
- Erweiterung der RVK-Online XML
 - Vereinheitlichung der impliziten Verweise
 - Verweise explizit als XML Entities
- Verbesserung des heuristischen Ansatzes
 - Ermitteln des gemeinsamen Nenners durch Aufsummieren der Ergebnisse „nach oben“ im RVK-Baum
 - Ermitteln von Notationspaaren mit mangelnder Trennschärfe
- Erweiterung des Vergleichs der Bezeichnungen
 - Betrachten von Teilpfaden
 - Betrachten von Teilähnlichkeiten





Vielen Dank für Ihre Aufmerksamkeit

<http://blog.bib.uni-mannheim.de/Classification>





Tabelle Titel mit RVK-Notationen

1	1805182
2	508173
3	160288
4	52055
5	16725
6	5768
7	2691
8	1578
9	676
10	776
11	56
12	11

Äquivalenzklassen