

# Mehrwertdienste für das Information Retrieval: das Projekt IRM

Philipp Mayr, Peter Mutschke, Philipp Schaer, York Sure

12. Tagung der Deutschen ISKO  
(International Society for Knowledge Organization)  
19.Oktober 2009, Bonn

## Problematik bei der Suche

- Informationssuche ist heute für Endnutzer alltäglich aber mit Problemen behaftet
- Suche/Retrieval ist zentral für die Informationswissenschaft

Typisch ist die Unterschiedlichkeit (Heterogenität) der:

- Suchräume (Google vs. Google Scholar vs. WoS)
- Dokument- und Medientypen
- Informationsanbieter (kommerziell vs. frei)
- Erschließung (kontrollierte Vokabulare vs. frei)
- Terminologie
- Benutzer (Einsteiger vs. Information-Professional)

# Hintergrund

## Nutzerperspektive: Was will der Nutzer?

- Relevante & qualitative Dokumente (relevance ranking)
- Umfassende Suche: Dokumente aus anderen Fächern, Datenbanken (eine Datenbank ist nicht genug)
- Flexible Suchsysteme: alternative Suchstrategien und -techniken (Filterung)
- Einfach Suchen

## IR-Mehrwertdienste

IR-Mehrwertdienste sollen das Retrieval und die Retrieval-Ergebnisse verbessern

- Vorschlagen von kontrollierten Termen (Search Term Recommender - STR)
- Re-Ranking von Ergebnisdokumenten nach dem ersten Listen der Ergebnisse
  - Bradfordizing
  - Autorenzentralität

# Schwerpunkt des Projekts: Prototyping und Evaluation der Dienste

- Prototyping
- Retrievalevaluation auf der Basis qualitativer Relevanzbewertungen in Clef und KoMoHe:
  - Vergleich mit konventionellen Rankingmethoden
  - Plausibilitätstests innerhalb der per Bradfordizing oder Autorenzentralität erstellten Rankings
  - Vergleich zwischen Bradfordizing und Autorenzentralität, statistische Zusammenhänge
- Retrievaltests mit Benutzern (Online-Bewertungs-Tool)
- Usability-Tests

# Search Term Recommender (STR)

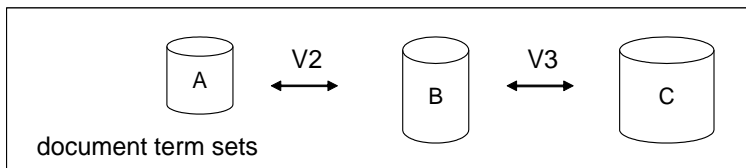
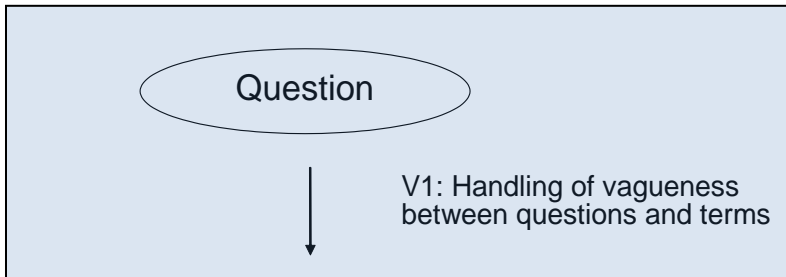
**Hintergrund: Sprachproblem im Information Retrieval  
(Abgleichen von Sprache)**

**Unterschiedl. Terminologien (Benutzer, Autor, Indexer)**

- Suchwortauswahl hat große Auswirkungen auf den Sucherfolg
- Vorschlagen von kontrollierten Termen (STR als Terminologiedienst) aus Co-occurrence Analysen (basierend auf Recommind Mindserver)
- Spezialisierte STR (Petras, 2006)

## Vocabulary Problem

- Furnas et al. (1987)
- Especially a problem in Digital Libraries



V2/V3:  
Bilateral handling of vagueness



## Co-Word analysis

### *Doc 1*

**Title:** Introduction to Information Retrieval

**Controlled Terms:** Information, retrieval

### *Doc 2*

**Title:** A survey of Information Retrieval and filtering methods

**Controlled Terms:** Information, retrieval

### *Doc 3*

**Title:** Expanding queries for the web

**Controlled Terms:** Information, retrieval, query expansion

### *Doc 4*

**Title:** Efficient and self-tuning incremental query expansion for top-k query processing

**Controlled Terms:** Query expansion

Co-word analysis by i.e.  
LSA, pLSA, SVM, ...

Term:

## Total hits: 2261

1. Prasad, Monica : *The Morality of Market Exchange: Love, Money, and Contractual Justice* (1999)
2. Cerulo, Massimo : *For Love or Money. The Commercialization of Intimate Life* (2008)
3. Ballinger, Lee : *In your face Sports for love and money* *Vor Deinen Augen Die Kommerzialisierung des Sports* (1981)
4. Onnen-Isemann, Corinna : *Money and Love. The Symbolic Significance of Money in Couple Relationships* (2004)
5. Nassehi, Armin : *Money and Love. Symbolic Significance of Money In a Couple's Relationship* (2005)
6. Stewart, B.K. : *The Regulation Of Sport Agents: For The Love Of Money* (1998)
7. Lopes Junior, Edmilson : *Love, Sex and Money: A Sociological Interpretation of the Market of Sexual Services* (2005)
8. Sawhill, Isabel; Thomas, Adam : *For Love and Money? The Impact of Family Structure on Family Income* (2005)
9. Binder, Amy : *For love and money: Organizations' creative responses to multiple environmental logics* (2007)
10. Bean, Frank D.; Van Hook, Jennifer; Brown, Susan K. : *For Love or Money? Welfare Reform and Immigrant Naturalization* (2006)

[next 10 documents >>](#)

[Economic Sociology](#) [Love](#)  
[Money](#) [Eroticism](#)  
[Dating \(Social\)](#)  
[Sexual Intercourse](#)  
[Opposite Sex](#)  
[Relations](#) [Interpersonal](#)  
[Relations](#) [Sexual Behavior](#) [Sexual](#)  
[Inequality](#) [Economic Theories](#)  
[Marital Relations](#)

- [Economic Sociology](#) [0.9582742]
- [Public Sphere](#) [0.42437306]
- [Love](#) [1.0]
- [Value \(Economics\)](#) [0.8274064]
- [Individual Differences](#)

## Re-Ranking

- Problem: klassische text-orientierte (tf-idf) und Web-IR Rankingverfahren sind nicht ideal für Retrieval in Digital Libraries (DL)
  - Dokumente mit wenig Textinformation
  - Volltext vs. Titelnachweise ohne Abstract
  - Zitationen fehlen weitestgehend (Ausnahme WoS)
- Ansatz: Re-Ranking soll Strukturen der Dokumentenräume der DL ausnutzen
  - Bradfordizing
  - Autorennetzwerke bzw. Autorenzentralität

# Motivation Bradfordizing

- Große Dokumentmengen für fachliche Suchanfragen sollen konzentriert werden
- hohe Robustheit der Bradford-Verteilung in Fachdatenbanken
- Plausibilität, dass der Nukleus (Kern) einer Bibliographie einen Nutzen erbringt
- Übertragbarkeit auf andere Dokumenttypen (z.B. Monographien)

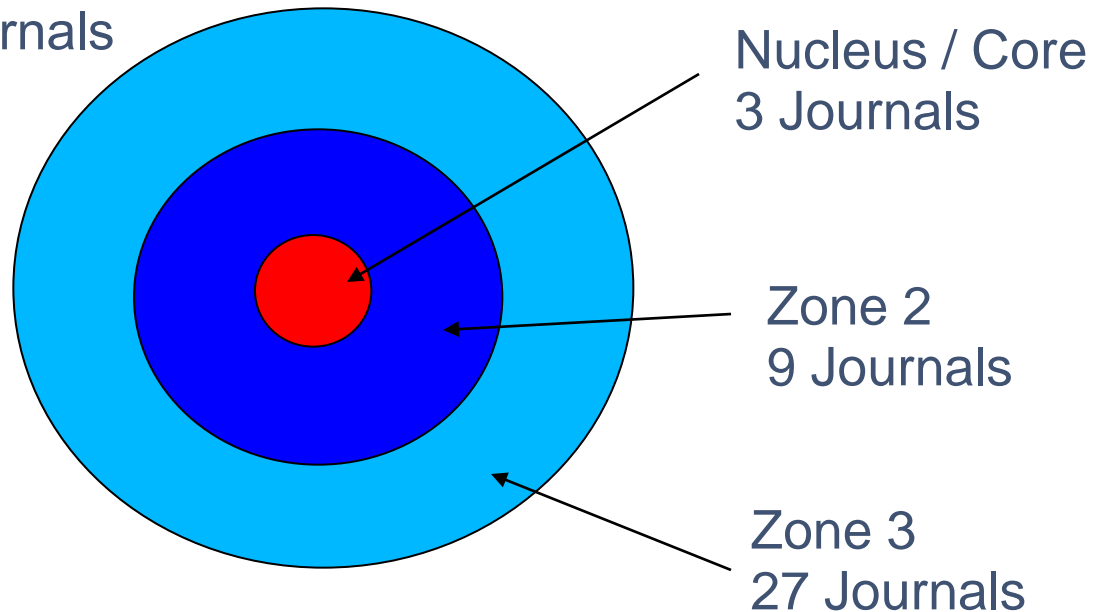
# Bradford Law Beispiel

Idealisiertes Beispiel: 450 Zeitschriftenartikel zu dem Topic „Systems theory“

150 Papers – 3 Journals

150 Papers – 9 Journals

150 Papers – 27 Journals



## Ansatz Bradfordizing

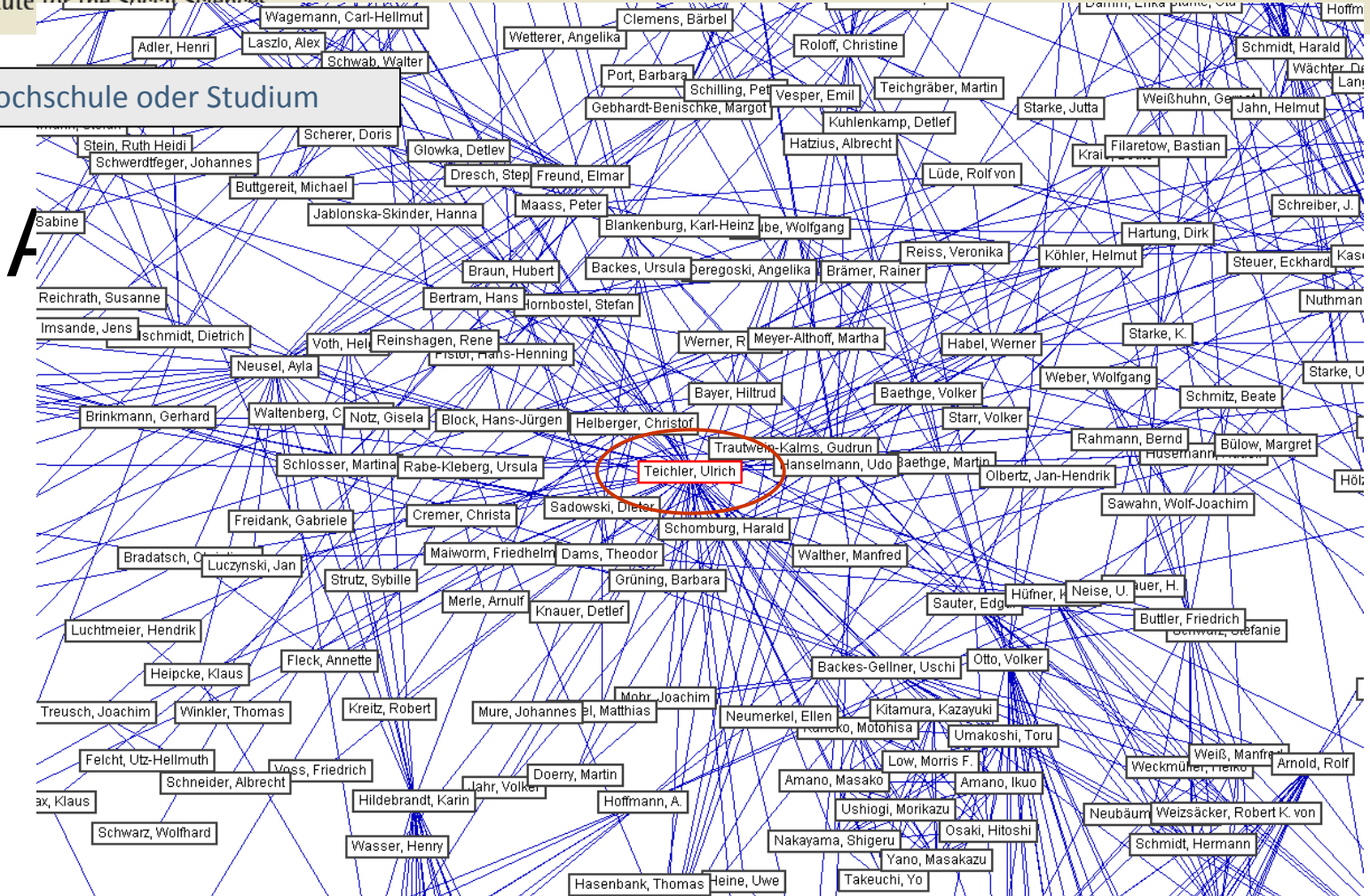
- Anwendung der bibliometrischen Methode Bradfordizing (White, 1981) für das Information Retrieval (IR); alternatives Re-Ranking von Ergebnismengen
  - Kernbereichsbildung (Kernliteratur) für Fachrecherchen (Kompensation, Strukturierung)
  - Kernzeitschriften und “Kernverlage” als Mehrwert
- Informetrics und IR

## Autorennetzwerke

- Motivation:
  - „Veröffentlichungen bekannter Autoren werden als relevanter eingeschätzt“ [Nutzerstudie Institut für Marketingforschung (2007): Erwartungen an wissenschaftliche Fachportale]
  - hohe Treffermengen -> höhere Anforderungen ans Ranking -> Nutzung von Informationen über die Struktur der Community
  - Expertensuchen
  
- Ansatz:
  - (sozialen) Status der Akteure innerhalb der soziale Netzwerke der Community evaluieren
  - Ranking nach dem
    - Grad der Integration der Akteure in den sozialen Netzwerken
    - Grad des Beitrages der Akteure zur Kohäsion des Netzwerkes

Schlagwort = Hochschule oder Studium

8304 SOLIS-Nachweise  
5172 vernetzte Akteure  
(1173 Giant)



Knoten = Autoren  
Kanten = Koautorenschaften  
Komponenten: maximal verbundener Teilgraph

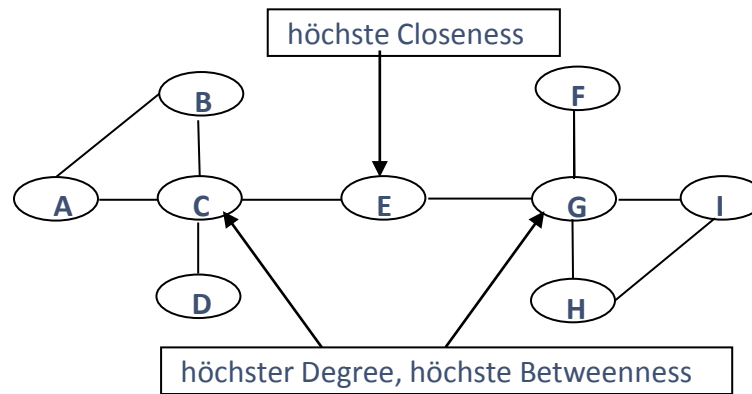
# Methode: Graphentheorie + Netzwerkanalyse

Zentralität von Autoren in Ko-Autorennetzwerken

- Dokumentranking nach dem Zentralitätswert der Autoren  
(Annotation Dokument mit dem maximalen Zentralitätswert der Autoren)

Zentralitätsmaße:

- **Degree:** Zahl der direkten Ko-Autoren (soziale Aktivität)
- **Closeness:** Nähe zu anderen Akteuren im Netzwerk (strukturelle Unabhängigkeit)
- **Betweenness:** Zahl der durch einen Akteur verbundenen Akteure (soziale Kontrolle)



# Netzwerkeffekte von Zentralität

Zentralitätsmaße messen (nur) das Einflusspotential eines Akteurs im Netzwerk

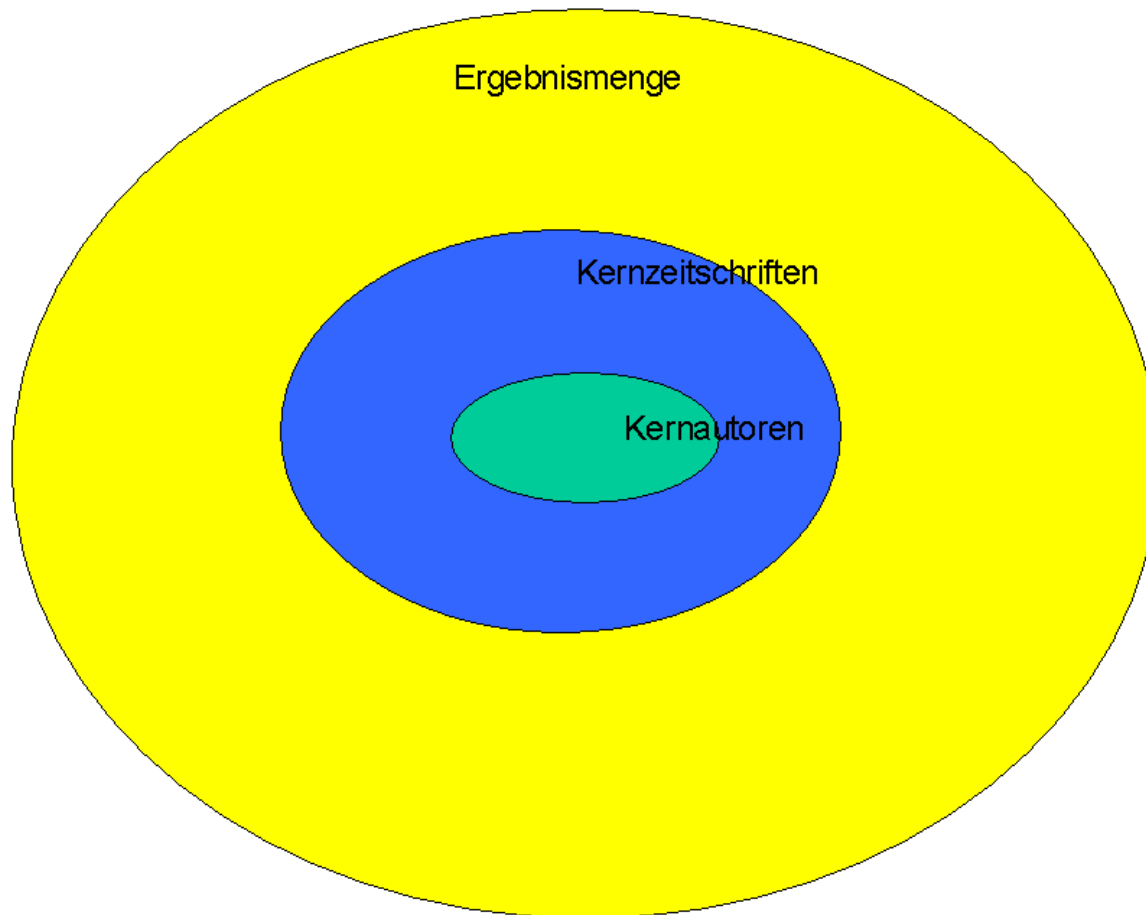
Frage: Welche Effekte haben zentrale Positionen auf Netzwerkprozesse?

Netzwerkforschung:

- Zusammenhang zwischen der Nennung von Führungspersonen und deren Zentralität in Kommunikationsnetzwerken (Bavelas 1950, Freeman et al. 1979/80)
- Zusammenhang zwischen der Zentralität von Teammitgliedern und deren Erfolg (Sparrowe et al. 2001)
- Zusammenhang zwischen zentraler Positionierung von Autoren in Ko-Autorennetzwerken und der Zentralität der von diesen Autoren behandelten Themen in Themennetzwerken (Mutschke/Qua-Haase 2001)

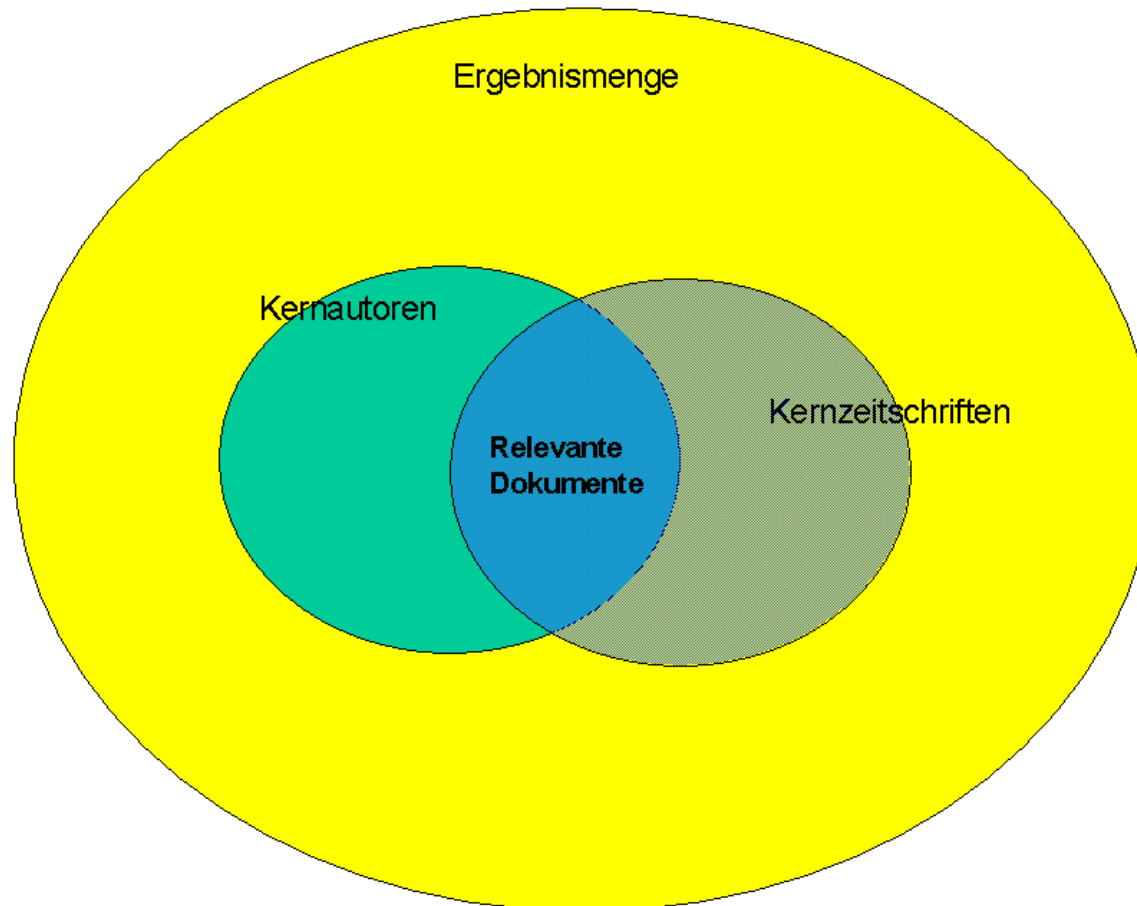
# Szenarios für kombinierte Mehrwertdienste

iterative Anwendung von Mehrwertdiensten auf die Ergebnismenge

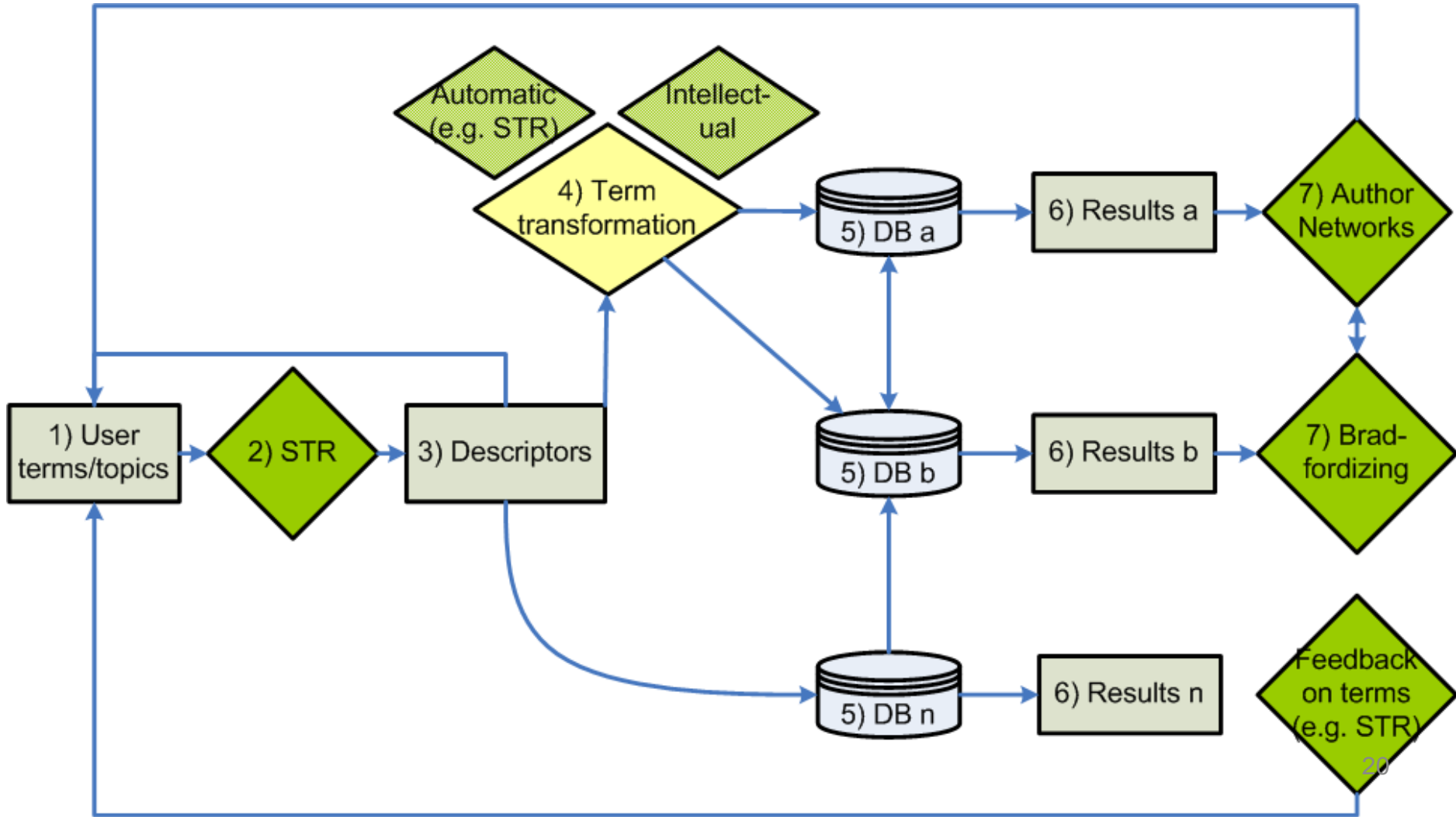


# Szenarios für kombinierte Mehrwertdienste

simultane Anwendung von Mehrwertdiensten auf die Ergebnismenge



## Zusammenspiel der Mehrwertdienste



# Evaluation

## Forschungsfragen

1. Verbessert sich das Retrieval durch Einsatz der Mehrwertdienste?
2. Werden durch die Mehrwertdienste ANDERE relevante Treffer favorisiert als durch Standardverfahren? Wenn ja, wie hoch ist der (quantitative) Mehrwert?
3. Werden die Mehrwertdienste vom Benutzer verstanden und angenommen (Akzeptanz)?
4. Gibt es eine (weitere) Verbesserung durch Kombination der Dienste? Wenn ja, welche Kombinationen sind am erfolgreichsten?

## Wer evaluiert

- User
  1. Interessierte Öffentlichkeit (z.B. sowiport-Nutzer)
  2. Informationsexperten (Dokumentare, Bibliothekare)
  3. Fachwissenschaftler (Soz., Sport, Bildung)
  4. Studenten-Gruppen (Koblenz, Berlin, ...)
    - Anfänger
    - Hauptstudium

→ Pro Gruppe und Disziplin mind. 20-30 Pers.

## Was wird evaluiert

- Dokumente (gerankt)
- Varianten von Trefferlisten der Services

### → Trefferlisten

- Klassisch: Doc-Listen
- Aggregiert je nach Service (Journal-Listen, Autorenlisten, Term-Listen)

# Zusammenfassung

- IR Mehrwertdienste können in unterschiedlichen Domänen, Datenbanken und Dokumenttypen erfolgreich angewendet werden.
- Einzelverfahren (modellgestützt) sind für sich sehr vielversprechend
- Der Mehrwert kann empirisch z.B. im Retrieval-Test am Beispiel signifikanter Precision-Verbesserungen verdeutlicht werden.
- Nutzer sind intuitiv zufrieden.
- Hohe Plausibilität und Nützlichkeit der Methode

## IRM Project

<http://www.gesis.org/index.php?id=2479>

E-mail: [philipp.mayr@gesis.org](mailto:philipp.mayr@gesis.org)

## Szenarios für kombinierte Mehrwertdienste

### Kombinationsmöglichkeiten der Mehrwertdienste

<b>Kombination</b>	<b>Autorennetzwerke</b>	<b>Bradfordizing</b>	<b>STR</b>
0	-	-	-
1	1	-	-
2	-	1	-
3	-	-	1
4	1	2	-
5	1	1	-
6	2	1	-
7	2	-	1
8	-	2	1
9	2	3	1
10	3	2	1
11	2	2	1

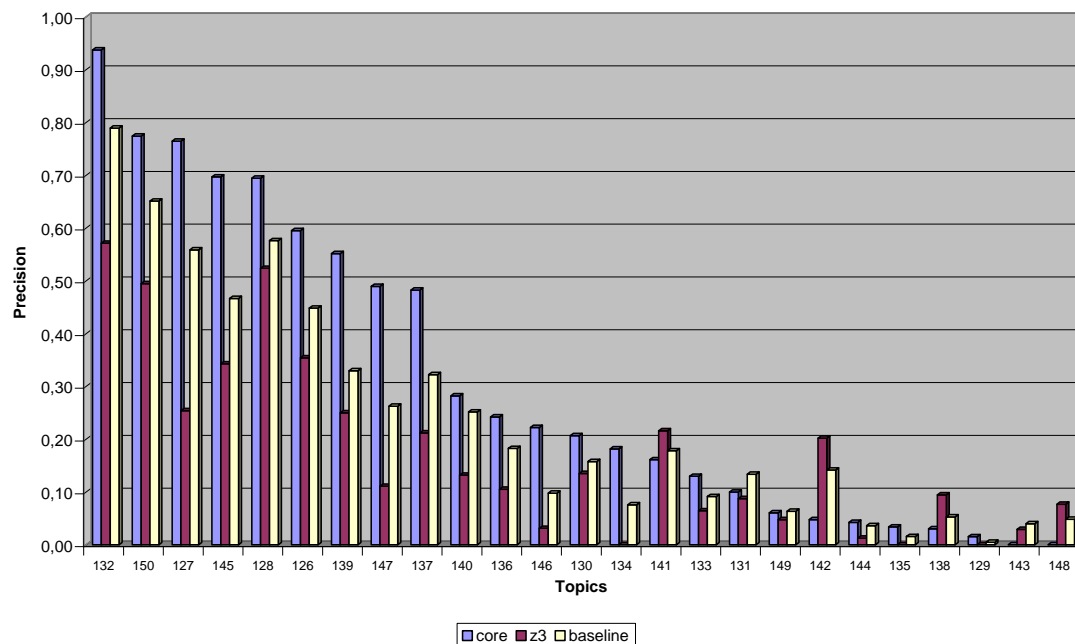
Zahl bedeutet Reihenfolge des Einsatzes.

(Legende: 1. Zeile (0) = Baseline, - = kein Einsatz).

Die gleiche Zahl bedeutet simultaner Einsatz mit Ergebnis-Merging.

## Verbesserung der Precision durch Bradfordizing (Mayr, 2009)

Articles for 25 topic (CLEF 2005)	Precision	Improvement
all articles (baseline)	0.239	
articles in core	0.310	(29.52%)*
articles in z3	0.174	
articles in core CLEF 2005 articles	0.310	(78.03%)*



## Ältere Evaluationsergebnisse

- Hohe Korrelation zwischen Autorenzentralität in Autorennetzwerken und der Zentralität der von diesen Autoren behandelten Themen in Themennetzwerken (Mutschke/Quan-Haase 2001)
- Positive qualitative Evaluation mit DJI 1997 und 2007
- Positiver Retrievaltest (Mutschke 2004):

Query	Ergebnismenge sortiert nach			Information. Mehrwert
	<i>PY</i>	<i>IDF</i>	<i>ACL</i>	<i>ACL (%)</i>
Jugend – Gewalt	0.25	0.60	0.55	92
Rechtsextremismus – Ostdeutschland	0.35	0.45	0.60	122
Frau – Personalpolitik	0.35	0.60	0.65	100
Widerstand – Drittes Reich	0.40	0.65	0.95	138
Zwangsarbeit – II. Weltkrieg	0.55	0.65	0.70	92
Eliten – BRD	0.40	0.70	0.85	107
Armut – Stadt	0.30	0.35	0.55	157
Arbeiterbewegung – 19./20. Jahrh.	0.55	0.55	0.90	164
Wertewandel – Jugend	0.40	0.50	0.30	50
Terrorismus - Demokratie	0.20	0.35	0.60	129
<b>Durchschnitt</b>	<b>0.38</b>	<b>0.54</b>	<b>0.67</b>	<b>115</b>