Ulrike Junger

Kann man Inhaltserschließung automatisieren?

Das Projekt PETRUS an der Deutschen Nationalbibliothek

DEUTSCHE NATIONAL BIBLIOTHEK

Kurzer Abriß

1. **Petrus – warum, was, wie?**
2. **Automatische Sachgruppenvergabe**
3. **Automatische Beschlagwortung**

Petrus - warum, was, wie

Die Millionenfrage:

Wie schafft man es, von einem Erschließungsvolumen von ca. 250.000 Medieneinheiten p.a. auf vielleicht 1.000.000 zu kommen?

Petrus - warum, was, wie

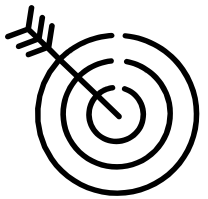
- Die Antwort:
  Petrus = Prozessunterstützende Software für die digitale Deutsche Nationalbibliothek

Petrus - warum, was, wie

Das Ziel:

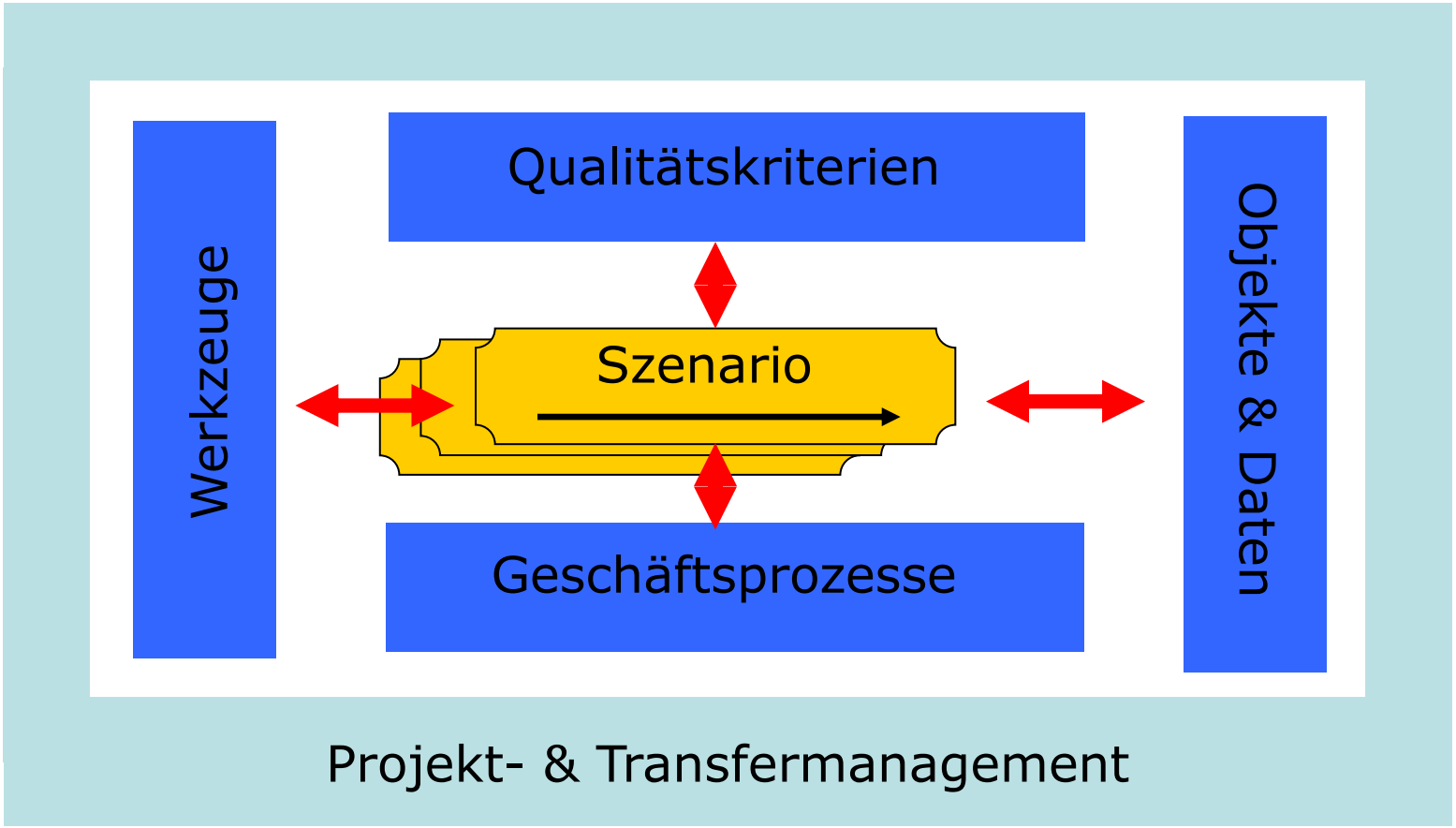
Effiziente Unterstützung des Erschließungsprozesses durch softwaregestützte Verfahren für

- Metadatenextraktion und –erzeugung
- Verknüpfung mit Normdaten
- verbale Erschließung
- Klassifikatorische Erschließung



Petrus - warum, was, wie

Arbeitspakete und Szenarien



Petrus - warum, was, wie

4 Szenarien

- Parallelausgaben
- Normdatenverknüpfung
- Automatische Sachgruppenvergabe
- Automatische Beschlagwortung

Petrus - warum, was, wie

Ausschreibungsverfahren „Erprobung softwaregestützter Verfahren für eine automatische Verschlagwortung und Klassifizierung“

Averbis, Freiburg Averbis Extraction Platform	Intrafind, Planegg bei München Intrafind TopicFinder
iSquare, Berlin iSquare SmartSearch	Rapid-i, Dortmund RapidMiner

Automatische Sachgruppenvergabe

Ziel:

- automatische Zuordnung zu einer Hauptsachgruppe
- Option: Vergabe von bis zu 2 Sachgruppen (Steuerung über Konfidenzwert)

Qualitätsziel:

- Richtige Zuordnung der Hauptsachgruppe in mindestens 80 % der Fälle
- Vergleichsmaß = intellektuell vergebene Sachgruppe.

Automatische Sachgruppenvergabe

- DNB-Sachgruppen: 101 Sachgruppen
- System basiert auf der Dewey-Dezimalklassifikation (DDC)
- Verschiedene Anwendungszwecke, primär Gliederung von Nationalbibliografie

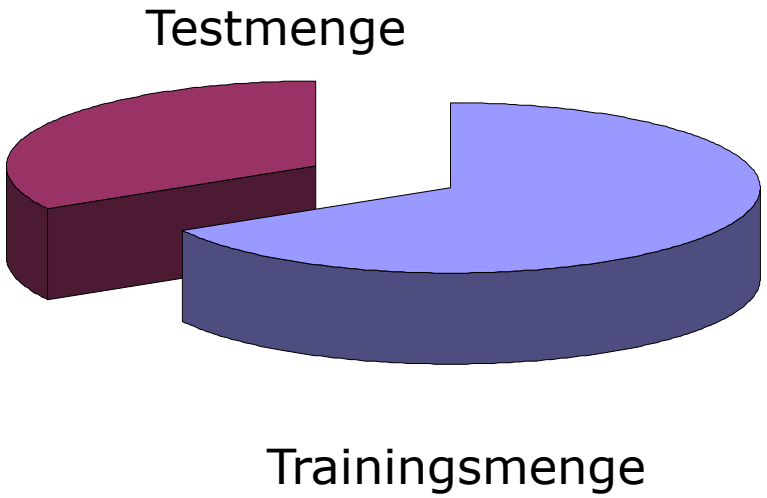
000 Allgemeines
004 Informatik
010 Bibliographien
…

Automatische Sachgruppenvergabe

Trainings-/Testkorpora

gescannte Inhaltsverzeichnisse (deutsch, DNB + HEBIS):

Umfang: ca. 120.000 Objekte



Volltexte (deutsch):

Online-Hochschulschriften
+ Online-Monografien

Umfang: ca. 45.000 Objekte

Automatische Sachgruppenvergabe
Herausforderungen und Schwierigkeiten

- ungleichmäßige Verteilung der Objekte auf die Sachgruppen
- Sachgruppen, die sich nicht gut abgrenzen lassen
- starke Heterogenität der Objekte
- Fehler bei der Formatkonvertierung

Automatische Sachgruppenvergabe

Testfall „Kreuztest“

Trainingskorpus: ca. 113.000 *Inhaltsverzeichnisse* aus 81 Sachgruppen

Testkorpus: ca. 46.500 *Volltexte* aus 45 Sachgruppen

In der Auswertung: bis zu **drei** Sachgruppen mit dem höchsten Konfidenzwert

Automatische Sachgruppenvergabe

Qualitätsmaße

Recall:

Anzahl richtig kategorisierter Objekte **/** Gesamtzahl der Objekte

Welcher Anteil der Objekte ist tatsächlich in der richtigen Kategorie gelandet?

Precision:

Anzahl richtig kategorisierter Objekte **/** Gesamtzahl der Zuordnungen zu dieser Kategorie

Welcher Anteil in einer Kategorie gehört tatsächlich dorthin?

F-Measure (Harmonisches Mittel):

2 * Precision * Recall **/** (Precision + Recall).

Automatische Sachgruppenvergabe Ergebnisse Testfall Kreuztest

DEUTSCHE NATIONAL BIBLIOTHEK

	Syst. A	System B	System C	System D
F-Measure$_1$*	0,75	0,68	0,71	0,65
Recall$_1$*	0,75	0,70	0,68	0,63
Precision$_1$*	0,77	0,70	0,76	0,86

	Syst. A	System B	System C	System D
F-Measure$_2$*	0,61	0,54	0,58	0,55
Recall$_2$*	0,88	0,83	0,85	0,79
Precision$_2$*	0,50	0,43	0,46	0,45

	Syst. A	System B	System C	System D
F-Measure$_3$*	0,54	0,48	0,53	0,51
Recall$_3$*	0,91	0,87	0,90	0,86
Precision$_3$*	0,43	0,37	0,41	0,40

*Gewichtetes Mittel über alle SG

Automatische Sachgruppenvergabe
Top Ten-Sachgruppen Testfall Kreuztest

DEUTSCHE NATIONAL BIBLIOTHEK

	System A			System B			System C			System D		
	SG	Anzahl	F-M.	SG	Anzahl	F-M.	SG	Anzahl	F-M.	SG	Anzahl	F-M.
1	**610**	20648	0,91	**610**	20556	0,89	**610**	20649	0,84	**610**	20649	0,83
2	**630**	2452	0,77	**780**	83	0,80	**780**	85	0,75	**780**	85	0,71
3	**510**	520	0,73	**510**	519	0,73	**630**	2452	0,71	**791**	63	0,69
4	**540**	2453	0,73	**340**	805	0,67	**230**	152	0,69	**340**	838	0,67
5	**780**	85	0,73	**540**	2448	0,67	**340**	838	0,69	**530**	2004	0,65
6	**230**	152	0,72	**100**	189	0,66	**540**	2453	0,69	**320**	337	0,63
7	**100**	189	0,70	**230**	149	0,66	**530**	2004	0,68	**540**	2453	0,63
8	**370**	936	0,70	**370**	929	0,64	**550**	569	0,67	**370**	936	0,61
9	**791**	63	0,70	**791**	63	0,64	**830**	238	0,67	**830**	238	0,61
10	**320**	337	0,68	**720**	184	0,63	**370**	936	0,65	**230**	152	0,6

Automatische Sachgruppenvergabe
Vorläufiges Fazit und weitere Schritte

- Die Systeme ähneln sich in den Ergebnissen
- Trainingskorpus und Testkorpus können unterschiedlich sein
- Das Ziel, eine Sachgruppe zu vergeben, wird beibehalten
- Aufbau eines neuen Testkorpus mit Volltexten aus der Reihe O und weitere Tests

Automatische Beschlagwortung
Ziele

Automatische Anreicherung von Online-Publikationen mit

- Schlagwörtern aus der SWD (= primäres Ziel)
- Schlagwörtern aus weiteren kontrollierten Vokabularen (optional)
- freien Deskriptoren (als weitere Sucheinstiege und zur Unterstützung der SWD-Pflege)

Automatische Beschlagwortung
Erster Testlauf

- Für die erste Stufe nur Integration der SWD-Sachschlagwörter (ca. 160.000)
- 12 ausgewählte Sachgruppen aus dem Bereich Sozialwissenschaften/STM
- Auswertung über Stichproben: intellektuelle Bewertung der automatisch vergebenen SWD-Schlagwörter

Bewertung der automatisch vergebenen SWD-Schlagwörter

automatisch vergebene SWD-SW	ID	sehr nützlich	nützlich	wenig nützlich	falsch/schädlich
Qualitätsmanagement	042190576	☐	☐	☐	☐
Standardisierung	040569144	☐	☐	☐	☐
Qualität	040479668	☐	☐	☐	☐
An-2	970390300	☐	☐	☐	☐
Bild	040065685	☐	☐	☐	☐
Normung	040426262	☐	☐	☐	☐
Modell	04039798X	☐	☐	☐	☐
Deutschunterricht	040119750	☐	☐	☐	☐

Automatische Beschlagwortung
Erster Testlauf

- Testkorpora:
  - ca. 42.000 Dokumente mit Inhaltsverzeichnissen
  - ca. 5.600 Volltexte
- Stichprobengröße: jeweils 16 Objekte pro Sachgruppe, Testkorpus und System

Automatische Beschlagwortung

0500 Aa
0600 ra;di;si
1100 2009
1500 /1ger
1700 /1XA-DE-HE
2000 978-3-8288-9907-0*kart. : EUR 29.90 (DE), EUR 29.90 (AT), sfr 52.20 (freier Pr.)
2040 9783828899070
2100 09,N18,0177
2105 09,A26,0362
2150 3282913
2230 Best.-Nr. 9907
3000 !138141525!*Bro´nska, Justyna*
4000 Die @Schweiz in Europa: mittendrin, doch außen vor? : Auswirkungen eines EU-Beitritts im Kontext der Erfahrungen Österreichs / Justyna Bro´nska
4030 Marburg : Tectum-Verl. ***5104661
4060 371 S.
4062 21 cm
4204 Zugl.: Bonn, Univ., Diss., 2008
4700 man
4701 hek
4715 =u $=c 04=d DNB=e 1
5050 320

Automatische Beschlagwortung

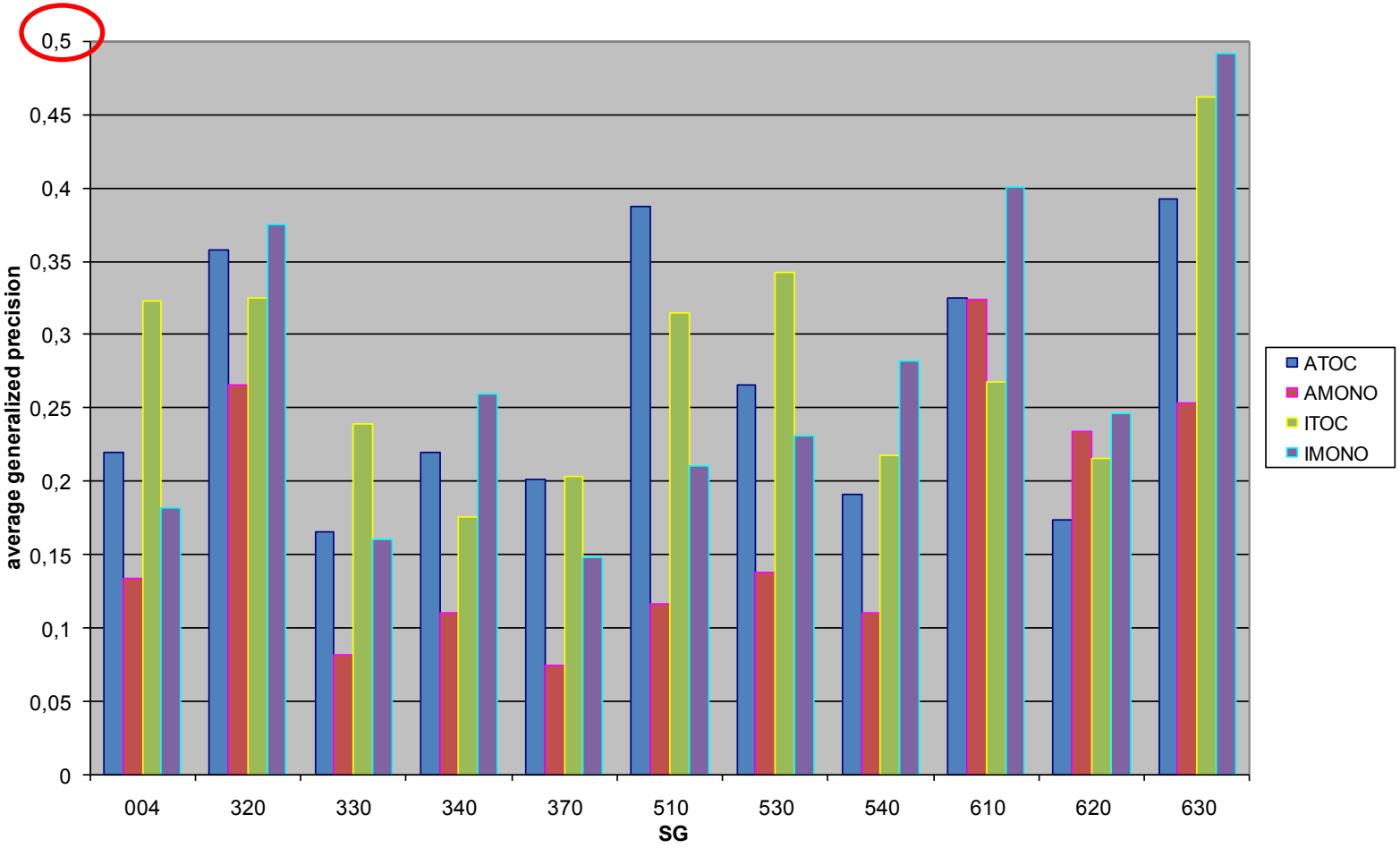
automatisch vergebene SWD-SW	ID	sehr nützlich	nützlich	wenig nützlich	falsch/schädlich
Schweiz	4713011-8	☐	☐	☐	☑
Volksabstimmung	4134790-0	☑	☐	☐	☐
Beitritt	4120988-6	☑	☐	☐	☐
Mitwirkung	4140386-1	☐	☑	☐	☐
Volksbegehren	4063810-8	☑	☐	☐	☐
Zeitungsartikel	4125430-2	☐	☐	☑	☐
Volksrechte	4188544-2	☐	☑	☐	☐
Eidgenossenschaft	4151157-8	☐	☑	☐	☐
Res	4552375-7	☐	☐	☐	☑
Marschroute	4750587-4	☐	☐	☐	☑
Monozyten-Makrophagen-System	4177897-2	☐	☐	☐	☑
direkteDemokratie	4134792-4	☑	☐	☐	☐
Demokratie	4011413-2	☐	☑	☐	☐
Schweizer	4643846-4	☐	☑	☐	☐
wirtschaftlicheIntegration	4066410-7	☐	☑	☐	☐
Projekt	4115645-6	☐	☐	☑	☐
europ?ischeIntegration	4071013-0	☑	☐	☐	☐
Parlament	4044685-2	☐	☑	☐	☐
daspolitische	4136977-4	☐	☐	☑	☐
Integration	4027238-2	☐	☐	☑	☐
Mitgliedschaft	4135885-5	☐	☐	☑	☐

Automatische Beschlagwortung

005 Ts1
011 /s
021 4713011-8
800 |s|Schweiz <Geographischer Name>
808 |a|Vorlage
808 |z|B59
809 |x|fo *erl
810 11.3b;19.1c
816 910.014#2# [2007-01-01]
850 |s|Geographischer Name
903 |e|DE-384
903 |r|DE-384

Automatische Beschlagwortung

Ergebnisse des ersten Tests



Automatische Beschlagwortung
Wie geht es weiter?

Verbesserung der Modelle

- Berücksichtigung klassifikatorischer Information
- Auswertung von Korrelationen zwischen Schlagwörtern
- Integration weiterer Schlagwortkategorien in den Thesaurus der Systeme

Laufender Test

- Stichproben von je 30 Titeln aus 16 Sachgruppen
- Vergabe freier Deskriptoren durch eines der Systeme
- Einschätzung der Inter-Bewerter-Konsistenz

Und sonst …

gibt es noch einige Baustellen, z.B.

- Gestaltung der Geschäftsprozesse
- Entwicklung eines Qualitätsmanagements
- Umgang mit den automatisiert erzeugten Daten im DNB-Portal und in den Datenlieferungen
- endgültige Systembeschaffung

DEUTSCHE NATIONAL BIBLIOTHEK

Vielen Dank für Ihre

Aufmerksamkeit!

Herzlichen Dank auch an meine Kollegen in der DNB für die Überlassung von Folien für den Vortrag, namentlich Christa Schöning-Walter!

Literaturhinweis:
Christa Schöning-Walter: PETRUS. In: Dialog mit Bibliotheken 2010/1, S. 15ff.

Kontakt:
Ulrike Junger
u.junger@d-nb.de