38. Deutscher Bibliothekartag, Erfurt

Fundy-Nationalpark
ul, 25. Mai 2008

VZG Projekt Colibri/DDC

Für eilige Leser

Ulrike Reiner

Automatische DDC-Klassifizierung von bibliografischen Titeldatensätzen

Version: Lite

VZG

Automatische DDC-Klassifizierung von bibliografischen Titeldatensätzen

- Inhalt des Vortrages -

Fundy-Nationalpark
ul, 25. Mai 2008

VZG Projekt Colibri/DDC

DDC System, Bibliografische Titeldatensätze

OCLC classify

VZG Colibri / DDC (Klassifizierungskomponente, Wettbewerb, Ergebnisse und Bewertung)

Colibri / DDC - Forschungsfrage *Q1*

Fundy-Nationalpark
ul, 25. Mai 2008

Ist es möglich, eine inhaltlich stimmige DDC-Titelklassifikation aller GVK-PLUS-Titeldatensätze automatisch zu erzielen?

Dewey-Dezimalklassifikation (DDC)
Mini-DDC-Ausschnitt


Fundy-Nationalpark
ul, 25. Mai 2008

Ebene

1. **0**00 (Informatik, ...) ...
   | ...
2. ... **02**0 (Bibliotheks- und Informationswissen. ...) ...
   | ...
3. ... **006** (Spezielle Computerverfahren) ...
   | ...
   ...
   | ...
12. ... **025.302855741** (Einsatz von Dateiorganisation ...)

DDC-Klassifzierung: ein Thema (Kleidung)– mehrere Systemstellen !

Fundy-Nationalpark
ul, 25. Mai 2008

Kleidung wird z. B. abgehandelt in Psychologie, Ethnologie und Künste

DDC-Klassifzierung: ein Thema (Der Apfel) – welche Systemstelle?

Fundy-Nationalpark
ul, 25. Mai 2008

DNB-Titeldatensatz DNB984784829 im MAB2-Format

...
026 DNB984784829
...
037bger
...
331 Der Apfel
...
540aISBN 978-3-938793-62-6 Pp. : EUR 98.00
540aISBN 3-938793-62-7 Pp. : EUR 98.00
...
700 |100ÎDNB
705a□a110□c110□eDDC22ger

Intellektuelle DDC-Klassifizierung

Der Apfel: 110 (Metaphysik)

Fundy-Nationalpark
ul, 25. Mai 2008

Aleksandar Kellenberg:
Metaphysische Untersuchungen –
Erster Teil: **Der Apfel**

VZG Projekt Colibri/DDC

VZG

OCLC Classify
(an experimental classification web service)
Der Apfel {372.133}


Fundy-Nationalpark
ul, 25. Mai 2008

Jockweg, Bernd:
Der Apfel
DDC Klasse: 372.133 (100%)
(Unterrichtsmaterialien-Primärbildung)

Intellektuelle DDC-Klassifizierung

Mitt liv, min frihet : {297,...,920.72}

Fundy-Nationalpark
ul, 25. Mai 2008

Mitt liv, min frihet : en selvbiografi / Ayaan Hirsi Ali ;
oversatt av Poul Henrik Poulsen
DDC Classification: 297, 305.486, 305.486092, 305.48697, 305.486971092, 920.0092, 920.72

Beispiel aus: [http://www.onb.ac.at/events/files/rype.ppt], p. 8

OCLC Classify

Mein Leben, meine Freiheit : {324.2092}

Fundy-Nationalpark
ul, 25. Mai 2008

Hirsi Ali, Ayaan: Mein Leben, meine Freiheit
(Holdings: 1)
DDC Klasse: 324.2092 (100%) (Politiker--Biografien)

OCLC Classify

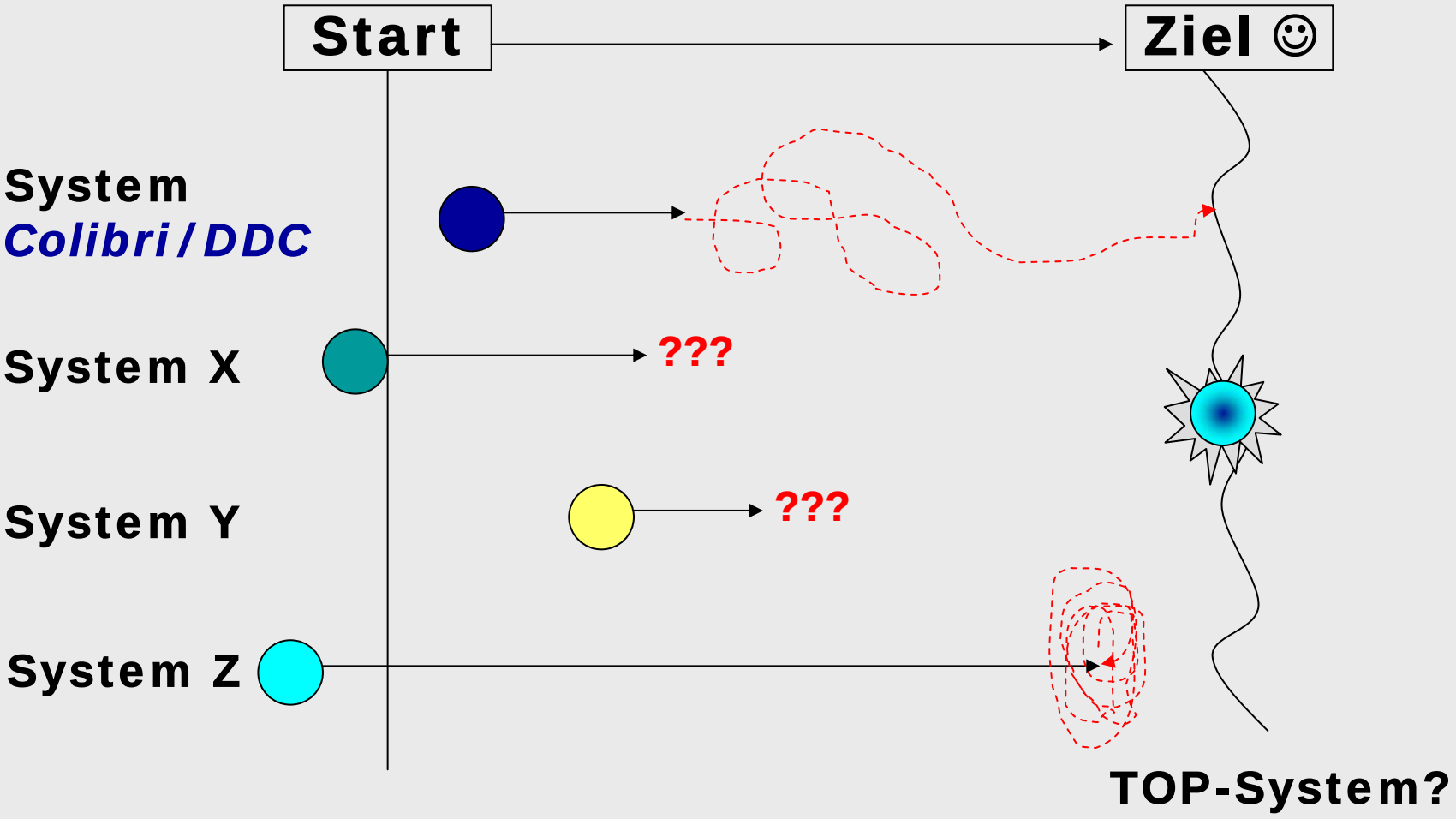
The caged virgin - an emancipation proclamation for women and Islam {297.082,...,922.97}

Fundy-Nationalpark
ul, 25. Mai 2008

Hirsi Ali, Ayaan: The caged virgin – an emancipation proclamation for women and Islam (Holdings: 1.407)
DDC Klassen: 297.082 (88.91%), 305.48697(6.68%), 922.97 (3.55%), Unclassified (4.41%)

Initiative Colibri/DDC-Wettbewerb (Juni 2009)
Ziel: bester automatischer DDC-Klassifizierer für bibliografische Titeldatensätze gesucht

Fundy-Nationalpark
ul, 25. Mai 2008



Initiative Colibri / DDC-Wettbewerb (Juni 2009)
Ziel: bester automatischer DDC-Klassifizierer für bibliografische Titeldatensätze gesucht

Fundy-Nationalpark
ul, 25. Mai 2008

VZG Projekt Colibri/DDC

Systemtest

Modell / Beschreibung des Systems
Zu testende Hypothesen
Bewertungskriterien und –maße
Daten ermitteln und bewerten

Automatische DDC-Klassifizierung (1)
Colibri/DDC-Modell (1)


Fundy-Nationalpark
ul, 25. Mai 2008

VZG Projekt Colibri/DDC

Pica+ : {..., 021A, ..., 044K, ...}; MAB2: {..., 310, ..., 410, ...}

{Apfel, Apfelbeere, Aronia}

634:= {..., <021A>-aronia, ...}

DNB991499077 := {..., <331>-aronia, <902s>-aronia, ...}

vc_DB := {000, 006.31, 025.302855741, 634 , ..., 999.23}

Automatische DDC-Klassifizierung (2)

Colibri / DDC-Modell (2)

Fundy-Nationalpark
ul, 25. Mai 2008

IR-Komponente

Vektorprodukt als Ähnlichkeitsmaß

DDC-Notationskandidat(en): DDC-Klasse mit größtem Ähnlichkeitswert

Automatische DDC-Klassifizierung (3)
Colibri/DDC-Modell (3)


Fundy-Nationalpark
ul, 25. Mai 2008

KI-Komponente

Heuristische Funktion

Obergrenze für Berücksichtigung von Häufigkeitswerten von Deskriptorwerten, Wert wird zur Laufzeit durch heuristische Regeln dynamisch bestimmt

Automatische DDC-Klassifizierung (4)
Colibri/DDC-Modell (4)


Fundy-Nationalpark
ul, 25. Mai 2008

KI-Komponente

Heuristische Regel, z. B. Berücksichtigung spezifischer Begriffe

Wenn Differenz zwischen zwei Häufigkeitswerten größer als „200“ **dann** *cutoff_val_dyn* := kleinerer Wert der beiden Häufigkeitswerte

Automatische DDC-Klassifizierung (5)
VZG Colibri / DDC-Suchsystem

Fundy-Nationalpark
ul, 25. Mai 2008

VZG Projekt Colibri/DDC

DDC-basierte Anfragen -> DDC-Suchsystem ->
DDC-basierte Antworten

Automatische DDC-Klassifizierung (6)
Softwaresystem-Architektur DDC-Suchsystem

Fundy-Nationalpark
ul, 25. Mai 2008

VZG Projekt Colibri/DDC

Eingabedaten -> DDC-Suchsystem mit Systemkomponenten (DDC-Klassifizierung, Frage-Beantwortung, DDC-Notationssynthese und –Analyse) -> Ausgabedaten

Automatische DDC-Klassifizierung (7)
Standard-Testbestände: Information Retrieval

Fundy-Nationalpark
ul, 25. Mai 2008

VZG Projekt Colibri/DDC

Cranfield, TREC, GOV2, CLEF, Reuters-21578, RCV1, 20 NEWSGROUPS

Automatische DDC-Klassifizierung (8)

Colibri/DDC-Systemtest: Testbestände

Fundy-Nationalpark
ul, 25. Mai 2008

DDC-Daten-/Wissensbasis 2004/2008
DNB-DDC-Testbestände 2007/2009

Andere Testbestände? Z. B.

100.000 BASE-Titeldatensätze
426.254 NORBOK-Titeldatensätze

Automatische DDC-Klassifizierung (9)

Colibri/DDC-Systemtest

Fundy-Nationalpark
ul, 25. Mai 2008

Kriterium für Tests/Experimente: Wiederholbarkeit!

Verwendete Hard- und Software: 2.2 GHz, 16GB Hauptspeicher. SuSE Linux Enterprise 10, gawk-3.1.5, ca. 1.250 Zeilen Programmcode.

Automatische DDC-Klassifizierung (10)
Eingabedaten


Fundy-Nationalpark
ul, 25. Mai 2008

DDC-System 2004: Elektronische Form als XML-Datei

DDC-Daten-Wissensbasis 2004: 3,0 Mio. Titeldatensätze

DDC-Daten-Wissensbasis 2008: 4,3 Mio. Titeldatensätze

DNB-DDC-Testbestand 2007: 25.653 Titeldatensätze

DNB-DDC-Testbestand 2009: 30.717 Titeldatensätze

Automatische DDC-Klassifizierung (11)
Datenkonvertierung (1)


Fundy-Nationalpark
ul, 25. Mai 2008

Datenkonvertierung

Eliminierung: irrelevante Deskriptorwerte, Sonderzeichen

Deskriptorwerte: Transliterierung, Kleinschreibung

Berücksichtigte MAB2-Felder

026, 037, 100, 310, 331, 335, 341, 370, 410, 412, 451, 540, 542, 700, 705, 902/12/22 s/g, 907/17/27 s/g

Automatische DDC-Klassifizierung (12)
Datenkonvertierung (2)


Fundy-Nationalpark
ul, 25. Mai 2008

Berücksichtigte Pica+ - Kategorien

001A, 003@, 004A, 004B, 004D, 005A, 006G, 006L, 006Y, 007G, 021A, 022A/01, 027D, 028A, 028B, 028C, 028E, 033A, 036C, 039B, 041A, 044A, 044C, 044E, 044F, 044G, 044K, 044L, 045A, 045F, 045Q, 045U,144Z/244Z, 145Z/245Z

Automatische DDC-Klassifizierung (13)

Fundy-Nationalpark
ul, 25. Mai 2008

Erstellung der (invertierten) DDC-Daten-/ Wissensbasis aus Daten des DDC-Systems und aus GVK-DDC-Titeldatensätzen

Größe der invertierten DDC-Daten-/Wissensbasis

2004 (510 MB): ca. 3 Min. Einlesezeit in den Hauptspeicher

2008 (712 MB): ca. 5 Min. Einlesezeit in den Hauptspeicher

Intellektuelle DDC-Klassifizierung

Der Apfel: 110 (Metaphysik)

Fundy-Nationalpark
ul, 25. Mai 2008

VZG Projekt Colibri/DDC

Aleksandar Kellenberg: **Metaphysische Untersuchungen** –
Erster Teil: **Der Apfel**

VZG

Automatische DDC-Klassifizierung (14)

Der Apfel: {334.683411, ... ,391.0092}

Fundy-Nationalpark
ul, 25. Mai 2008

correlation(dnb_A0745_DNB0984784829#ger#dno_i{110}#dno_a{M300,D330,S338,s334.683411,s338.108,s338.10942,s338.13,s338.17411,s343.73084,s391.0092}#consi: 2#matched: 1,1{apfel}): 00x.xxx xxx xxx xxx (0)

Automatische DDC-Klassifizierung (15)
?

Fundy-Nationalpark
ul, 25. Mai 2008

Petra Neumayer; Birgit Funfack: **Aronia – Powerbiostoffe aus der Apfelbeere**

Automatische DDC-Klassifizierung (16)

Powerbiostoffe aus der Apfelbeere : {615.321}

Fundy-Nationalpark
ul, 25. Mai 2008

correlation(dnb_A0912_DNB0991499077#ger#dno_i{615.32373}#dno_a{M600,D610,S61,s615.321}#consi:11#matched: 2,2{naturheilmittel, vitalitaet}):
111.110 00x xxx xxx (0.625)

VZG Projekt Colibri/DDC

VZG

Automatische DDC-Klassifizierung (17)
Bewertung (1)


Fundy-Nationalpark
ul, 25. Mai 2008

Projekt Colibri / DDC

Korrelationsmaße *C*, *CP* und *CN*

Stellenweiser Ziffernvergleich von links nach rechts zwischen intellektuell vergebener (*dno_i*) und automatisch ermittelter DDC-Notation (*dno_a*)

Automatische DDC-Klassifizierung (18)
Bewertung (2)


Fundy-Nationalpark
ul, 25. Mai 2008

Projekt Colibri / DDC

dno_i = 158.1 (Persönliche Weiterentwicklung und Analyse)

dno_a = 158 (Angewandte Psychologie)

CP = 111.0xx xxx xxx xxx; *CN* = (1+1+1+0)/4 = 0.75

Automatische DDC-Klassifizierung (19)
Bewertung (3)


Fundy-Nationalpark
ul, 25. Mai 2008

VZG Projekt Colibri/DDC

Projekt Scorpion / DDC

Bewertungsmaße H (=S1), T(=S2), O (=S3), G (=S4),
Sp (=S4), Co (=S5), Sy (=S6), B (=S7),
E (=S8), Cl (=S9), R (=S10)

Automatische DDC-Klassifizierung (20)
Bewertung (4)


Fundy-Nationalpark
ul, 25. Mai 2008

VZG Projekt Colibri/DDC

Projekt Pfeffer/RVK

Vergleich der automatischen und manuellen Klassifikation
Suche des nächsten gemeinsamen Vaterknoten im RVK-Baum

Bewertungsmaß: perfekt (=P1), gut (=P2),
mäßig (=P3), schlecht (=P3)

Automatische DDC-Klassifizierung (21)
Bewertung (5)


Fundy-Nationalpark
ul, 25. Mai 2008

Colibri/DDC					Scorpion/DDC						Pfeffer/RVK	
Bsp.dno_i	dno_a	CP	C	CN	S1	S2	S3	S4	S8	S9	S10	P
14. 111.8	110	110.0	2	0.50	x	x					x	P2
15. 571.5929	571	111.0000	3	0.43	x	x	x	Sp		x	x	P3
16. 111.85	110	110.00	2	0.40	x	x					x	P2

Vergleichende Betrachtung mit unterschiedlichen Bewertungsmaßen

Automatische DDC-Klassifizierung (22)
Bewertung (6)


Fundy-Nationalpark
ul, 25. Mai 2008

Klasse K	Mensch urteilt: (korrekte Zuordnung) gehört zu K	gehört nicht zu K
Maschine ermittelt: gehört zu K	a	b
Maschine ermittelt: gehört nicht zu K	c	d

Vierfeldertafel

Automatische DDC-Klassifizierung (23)
Bewertung (7)


Fundy-Nationalpark
ul, 25. Mai 2008

Bewertungsmasse

Precision P = a / (a+b)

Recall R = a / (a+c)

Fallout F = b / (b+d)

F-Measure = 2*P*R / (P+R)

Automatische DDC-Klassifizierung (24)
Bewertung (8)


Fundy-Nationalpark
ul, 25. Mai 2008

Weitere Bewertungsmasse

Accuracy, Error, Percent too specific, Percent too general, Average overlap, Accuracy at level, Eleven-point average precision, Precision-recall breakeven point

Automatische DDC-Klassifizierung (25)
Bewertung (9)

Fundy-Nationalpark
ul, 25. Mai 2008

DNB-Titeldatensätze mit intellektuell vergebener DDC-Notation -> DDC-Suchsystem -> DNB-Titeldatensätze mit intellektuell und automatisch vergebenen DDC-Notationen -> automatische Bewertung -> automatisch bewertete Klassifizierungsergebnisse

Automatische DDC-Klassifizierung (26)

Klassifizierungsergebnisse

Fundy-Nationalpark
ul, 25. Mai 2008

Name der Ergebnisdatei *res...*	res (Anz.)	tit (Anz.)	t [min]
res_vc_IDB-2004_in_DNB-2007	16.694	25.653	133
res_vc_IDB-2008_in_DNB-2007	15.365	25.653	136
res_vc_IDB-2004_in_DNB-2009	21.591	30.717	120
res_vc_IDB-2008_in_DNB-2009	21.422	30.717	140

Automatische DDC-Klassifizierung (27)
Automatisch bewertete Ergebnisse (1)


Fundy-Nationalpark
ul, 25. Mai 2008

Hypothesen

a) Unterschied bei unterschiedlichen Daten-/Wissens-basen ist signifikant

b) Unterschied bei verschiedenen Testbeständen ist nicht signifikant

Automatische DDC-Klassifizierung (28)

Automatisch bewertete Ergebnisse (2)

Fundy-Nationalpark
ul, 25. Mai 2008

Übereinstimmung mindestens in der DDC-Hauptklasse
63,85% (CN>0)

Verteilung der Übereinstimmungen, z. B.
36,15% (C=0); 13.71% (C=1); 26,29% (C=2)

Automatische DDC-Klassifizierung (29)

Automatisch bewertete Ergebnisse (3)

Fundy-Nationalpark
ul, 25. Mai 2008

Hypothese

Es gibt signifikante Unterschiede zwischen den DDC-Klassen (gemessen mit *CN*)

Automatische DDC-Klassifizierung (30)
Automatisch bewertete Ergebnisse (4)


Fundy-Nationalpark
ul, 25. Mai 2008

Hypothese

Es gibt signifikante Unterschiede
zwischen den DDC-Klassen
(gemessen mit Recall, Precision, F-Measure)

Automatische DDC-Klassifizierung (31)
Automatisch bewertete Ergebnisse (5)

Fundy-Nationalpark
ul, 25. Mai 2008

VZG Projekt Colibri/DDC

Hypothese

Es gibt keinen signifikanten Unterschied zwischen deutschen und englischen Titeldatensätzen

Automatische DDC-Klassifizierung (32)

Automatisch bewertete Ergebnisse (6)

Fundy-Nationalpark
ul, 25. Mai 2008

Hypothese

Es gibt signifikante Unterschiede zwischen den Reihen A, B und H

Automatische DDC-Klassifizierung (33)

Automatisch bewertete Ergebnisse (7)

Fundy-Nationalpark
ul, 25. Mai 2008

Hypothese

Es gibt signifikante Unterschiede hinsichtlich der Stelligkeit der DDC-Notationen

Stand der automatischen DDC-Klassifizierung

Fundy-Nationalpark
ul, 25. Mai 2008

Eingabedaten
einzelne - ggf. mehrere zusammenhängende – Wörter,
keine Volltexte, keine linguistischen Verfahren, kein Lexikon

Ermittlung der DDC-Notationskandidaten
Algorithmus verwendet IR- und KI-Verfahren

IR: Vektorprodukt, binäre Vektoren
KI: heuristische Regeln

2 Klassenaggregationen für **Ergebnisausgabe**

Automatische Bewertung der Klassifizierungsergebnisse

Perspektiven zur automatischen DDC-Klassifizierung (1)

Fundy-Nationalpark
ul, 25. Mai 2008

Vergrößerung der **DDC-Daten-/Wissensbasis**

Verbesserung der **Sacherschliessung** bei unzureichend erschlossenen Titeldatensätzen

Erweiterung der heuristischen Funktion, Verwendung weiterer **(KI/IR)-Algorithmen**, Lexikonerstellung

Eliminierung weiterer **irrelevante**r **Deskriptorwerte**

Andere Methode der (Klassenaggregation zur) **Ergebnisausgabe**

Anreiz durch **Colibri/DDC-Wettbewerb** ☺

Perspektiven zur automatischen DDC-Klassifizierung (2)

Fundy-Nationalpark
ul, 25. Mai 2008

Zuallererst neue Energie aufnehmen ...

Fundy Nationalpark, ul, 25. Mai 2008

Vielen Dank für Ihr Interesse am VZG-Projekt Colibri/DDC

Für Colibri-Unterstützung und –Diskussionen gilt mein Dank:
Dipl.-Kfm. Reiner Diedrichs, Direktor der Verbundzentrale (VZG) des GBV, Göttingen
Prof. Dr. Erhard Konrad (i.R.), Fakultät Elektrotechnik und Informatik, TU Berlin, Berlin
Dipl.-Inform. (FH) Alfred Vogelbacher, Network Support Engineer Solaris, Sun Microsystems GmbH, Berlin

Literatur:
Information Retrieval & Bewertung (1)


Fundy-Nationalpark
ul, 25. Mai 2008

[SALTON 1971] **The SMART Retrieval System – Experiments in Document Processing** (ed. Gerard Salton). Prentice-Hall, Englewood Cliffs, NJ, 1971.

[Jones 1981] Karen Spärck Jones: **Information Retrieval Experiment.** Butterworths, London, 1981.

[Jones 1996] Karen Spärck Jones; Julia R. Galliers: **Evaluating Natural Language Processing Systems. An Analysis and Review.** Lecture Notes in Artificial Intelligence 1083. Springer, Berlin, 1996.

[Voorhees/Harman 2005] **TREC: Experiment and Evaluation in Information Retrieval** (ed. by Ellen M. Voorhees; Donna K. Harman). MIT Press, Cambridge Massachusetts, 2005.

Literatur:
Information Retrieval & Bewertung (2)


Fundy-Nationalpark
ul, 25. Mai 2008

[Moens 2000] Marie-Francine Moens: **Automatic Indexing and Abstracting of Document Texts.** Kluwer Academic Publishers, London, 2000.

[Manning/Raghavan/Schütze 2008] Christopher D. Manning; Prabhakar Raghavan; Hinrich Schütze: **Introduction to Information Retrieval.** Cambridge University Press, Juli 2008. Online: http://www-csli.stanford.edu/~hinrich/information-retrieval-book.html.

[CLEF] **Cross-Language Evaluation Forum (CLEF)** . Online: http://www.clef-campaign.org/.

Literatur:
Automatische Klassifizierung & Bewertung (1)


Fundy-Nationalpark
ul, 25. Mai 2008

[Reiner 2008] Ulrike Reiner: **DDC-based Search in the Data of the German National Bibliography.** In: New Perspectives on Subject Indexing and Classification. Essays in Honour of Magda Heiner-Freiling. Deutsche Nationalbibliothek. Leipzig, Frankfurt am Main, Berlin, 2008, pp. 121-129.

[Reiner 2009] Ulrike Reiner: **Bewertung von automatisch DDC-klassifizierten Titeldatensätzen der Deutschen Nationalbibliothek (DNB).** VZG-Colibri-Bericht 1/2008. Online: http://taipan.dyndns.org/~ul/colibri05.pdf.

[Oberhauser 2004] Otto Oberhauser: **Automatisches Klassifizieren. Verfahren zur Erschliessung elektronischer Dokumente.** Master's Thesis. Zusatzstudiengang Bibliotheks- und Informationswissenschaft. Fakultät für Informations- und Kommunikationswissenschaften, Fachhochschule Köln, 2004.

Literatur:
Automatische Klassifizierung & Bewertung (2)


Fundy-Nationalpark
ul, 25. Mai 2008

[Wille 2006] Jens Wille: **Automatisches Klassifizieren bibliographischer Beschreibungsdaten - Vorgehensweise und Ergebnisse**. Diplomarbeit. Studiengang Bibliothekswesen Fakultät für Informations- und Kommunikationswissenschaften, Fachhochschule Köln, 2006.

[Pfeffer 2008] Magnus Pfeffer: **Automatische Vergabe von RVK-Notationen mittels fallbasiertem Schließen.** Vortrag: 97. Deutscher Bibliothekartag. 5. Juni 2008, Mannheim.

[Mehler/Waltinger 2009a] Alexander Mehler; Ulli Waltinger: **Automatic Enrichment of Metadata**. Vortrag: „9th International Bielefeld Conference". 4. Februar 2009, Bielefeld.

[Mehler/Waltinger 2009b] Alexander Mehler; Ulli Waltinger: **Enhancing Document Modeling by Means of Open Topic Models: Crossing the Frontier of Classification Schemes in Digital Libraries by Example of the DDC.** Wird publiziert in: Library Hi Tech, 2009.

Fundy-Nationalpark
ul, 25. Mai 2008

Für an Details interessierte Leser

Ulrike Reiner

Automatische DDC-Klassifizierung von bibliografischen Titeldatensätzen

Version: Full

http://www.opus-bayern.de/bib-info/volltexte/2009/736/