gesis
Leibniz-Institut für Sozialwissenschaften

Neue Verfahrenswege der Wissensorganisation

Eine Evaluation automatischer Indexierung in der sozialwissenschaftlichen Fachinformation

Dr. Andreas Oskar Kempf

13. Tagung der Deutschen ISKO, Potsdam, 19.-20.03.2013

Überblick

Inhaltserschließung in Zeiten der Automatisierung

Automatisierungsbemühungen in der sozialwissenschaftlichen Fachinformation (GESIS)

- Konzeptionelle Vorüberlegungen (Schalenmodell)
- Evaluationsdesign und Testeinstellungen
- Testergebnisse
  - auf Ebene der Dokumentstichprobe
  - auf Ebene der Einzeldokumente

Fazit und Ausblick

Automatisierungsbemühungen in der Inhaltserschließung

Erfahrungen mit Erprobung und Einsatz (semi)automatischer Erschließungsverfahren

- Bibliothekswelt u.a.:

  DNB

- Fachinformation u.a.:

  ZBW, ZPID, DIPF, GESIS

Unterschiedliche Ausgangsbedingungen und automatische Verfahrenstechniken → schwere Vergleichbarkeit

Konzeptuelle Vorüberlegungen (Schalenmodell)

- Differenzierung nach unterschiedlichen Niveaustufen der Datenrelevanz und Erschließungsqualität

  „Die innerste Schale enthält den Kern der relevanten Literatur. Er wird möglichst tief und qualitativ hochwertig erschlossen“ (Krause, 1996: 18).

- Vorschlag der automatischen Indexierung für nachgeordnete Schalen

  „Schale 4 enthielte die Ansetzung der Bibliotheken. Neben den gebundenen Deskriptoren (Beispiel Autor) steht für die Inhaltserschließung nur der Titel zur Verfügung, der automatisch indexiert wird.“ (ebd.: 19).

Evaluationsdesign

Berechnungsgrundlage bildete die intellektuelle Erschließung, d.h. berechnet wurde die Korrektheit der Deskriptor- und Notationszuordnungen:

Precision (Genauigkeit der Zuordnung) =

Anzahl der sowohl maschinell als auch intellektuell vergebenen SW. bzw. Notationen

Gesamtanzahl maschinell generierter SW bzw. Notationen

Recall (Vollständigkeit der Zuordnung) =

Anzahl der sowohl maschinell als auch intellektuell vergebenen SW bzw. Notationen

Gesamtzahl der intellektuell vergebenen SW bzw. Notationen

Indexierungskonsistenz (Grad an Übereinstimmung zw. masch. und intell. Erschließung) =

Anzahl der sowohl maschinell als auch intellektuell vergebenen SW bzw. Notationen

Gesamtzahl sämtlicher sowohl maschinell als auch intellektuell vergebenen SW bzw. Not.

Testeinstellungen

Testreihe I

Trainingsdaten: Gesamt-SOLIS (bis Juli 2009)

Testdaten: Dokumentstichprobe (280 Dokumente)

Testlauf I: Standardeinstellungen der MindServer-Software

Testlauf II: Erhöhung *Precision*,
Einführung eines *cut-off Levels*

Testreihe II

Trainingsdaten: Fachteilgebietsspezifische Untermengen (bis Juli 2009)

Testdaten: Fachteilgebietsspezifische Untermengen der Stichprobe (s.o.)

Testlauf III: Standardeinstellungen der MindServer-Software

Testlauf IV: Ausgeglichenes Verhältnis Recall/Precision

Ergebnisse Dokumentstichprobe (1/3)
Testläufe I und II

	Deskriptorvergabe		Notationsvergabe	
	TL I (n=271)	TL II (n=263)	TL I (n=262)	TL II (n=254)
Recall	44,0% (48,6%)	30,9% (36,4%)	63,0% (70,8%)	58,3% (79,2%)
Precision	32,4% (35,8%)	**53,0%** (62,3%)	37,5% (42,5%)	**40,0%** (54,3%)
Indexierungs-konsistenz	37,3% (41,3%)	39,0% (46,0%)	45,2% (52,0%)	46,4% (61,8%)

Erhöhung des Precision-Anteils und *Einführung eines cut-off Levels* führen zu einer **Erhöhung des Precision-Ergebnisses** sowohl für die SW- als auch für die Notationsvergabe

Ergebnisse Dokumentstichprobe (2/3)
Vergleich der Testläufe I - IV

Deskriptor-vergabe	TL I	TL II	TL III	TL IV
Soziologie (n=56)				
Recall	**46,6%**	*32,5%*	**45,1%**	42,1%
Precision	**31,5%**	*51,4%*	**37,4%**	41,1%
Ind.-Kons.	**37,6%**	*39,9%*	**40,9%**	41,6%
Politik (n=68)				
Recall	**46,0%**	*32,2%*	**43,5%**	41,5%
Precision	**36,1%**	*60,1%*	**42,4%**	46,4%
Ind.-Kons.	**40,0%**	*41,9%*	**42,9%**	43,2%
Geistesw. (n=40)				
Recall	**42,0%**	*30,0%*	**38,3%**	39,5%
Precision	**32,0%**	*51,4%*	**36,0%**	41,0%
Ind.-Kons.	**36,3%**	*37,9%*	**37,1%**	41,7%

Der Aufbau fachteilgebietsspezifischer MindServer-Versionen führt zu einer **Erhöhung von Precision** und **Indexierungskonsistenz** (vgl. TL I u. TL III).

Durch eine **stärkere Gewichtung der Precision** (Systemeinstellungen) und die **Einführung eines cut-off Levels** kommt die automatische Indexierung bei der SW-Vergabe der intellektuellen Indexierung am nächsten (siehe TL II).

Ergebnisse Dokumentstichprobe (3/3)
Vergleich der Testläufe I - IV

Notation	TL I	*TL II*	TL III	TL IV
Soziologie (n=56)				
Recall.	**56,5%**	*56,5%*	**56,5%**	39,1%
Precision	**30,2%**	*36,1%*	**33,3%**	47,4%
Ind.-Kons.	**40,5%**	*45,4%*	**42,0%**	54,3%
Hauptnot. gef.	**78,2%**	*78,4%*	**85,2**	69,2
Ranking-Position	**1,7**	*1,6*	**1,4**	1,2
Politik (n=68)				
Recall	**82,6%**	*73,9%*	**73,9%**	60,9%
Precision	**47,5%**	*47,2%*	**58,6%**	73,7%
Ind.-Kons.	**58,8%**	*58,3%*	**64,4%**	66,0%
Hauptklass. gef.	**92,4%**	*87,3%*	**89,4**	87,8
Ranking-Position	**1,7**	*1,5*	**1,2**	1,4
Geistesw. (n=40)				
Recall	**43,5%**	*43,5%*	**56,5%**	34,8%
Precision	**26,3%**	*30,3%*	**38,2%**	42,1%
Ind-Kons.	**33,9%**	*37,1%*	**44,8%**	39,5%
Hauptklass. gef.	**63,2%**	*59,0%*	**77,5**	57,9
Ranking-Position	**2,7**	*1,7*	**1,4**	1,1

Der Aufbau fachteilgebietsspezifischer MindServer-Versionen führt zu einem **höheren Ranking der intellektuell vergebenen Hauptnotation**.

Fachteilgebietsspezifische MindServer-Versionen *können* ebenso wie die Einführung eines cut-off Levels zu einer höheren Präzision bei der Notationsvergabe beitragen.

Ergebnisse Dokumentebene (1/1)

Deskriptorvergabe:

- Mangelnde Erkennung von Wortbindestrichtilgungen und adjektivisch gebrauchten Begriffen (Derivation)
- Mangelnde Grundformreduktion sowie Kompositazerlegung
- Semantische ‘Blindheit’ bei Negativ-Formulierungen und Auslassungen
- Gelungene Erkennung latenten Dokumentinhalts (z. B. Personennamen)
- Negative Auswirkungen der Kontexterkennung bei Dokumenten, die eine Vielzahl von Themen oder Forschungsperspektiven beinhalten
- Anfälligkeit der Kontexterkennung für irreführende Begriffsassoziationen

Notationsvergabe:

- (Inter)Disziplinäre Breite und Themenvielfalt können zu sehr vielen oder gar keinen Notationsvorschlägen führen
- Vermeintlich falsche Notationszuordnungen aufgrund der mangelnden Trennschärfe mancher Klassifikationsnotationen

Fazit und Ausblick

- Die automatisch generierten Indexierungsergebnisse weisen für die **Kerngebiete** der Datenbank SOLIS tendenziell eine **höhere Übereinstimmung mit dem intellektuellen Indexat** auf als für die Randgebiete – insbesondere bei der SW-Vergabe.
- Der **Aufbau fachteilgebietsspezifischer MindServer-Versionen** führt für die SW-Vergabe bei den untersuchten Fachteilgebieten zu einer **präzisieren Indexierungsleistung**.
- Die **Einführung eines cut-off Levels** führt zu einem **höheren Precision-Anstieg** als er durch den Aufbau fachteilgebietsspezifischer MindServer Versionen erzielt wird, d.h. bei Einführung eines semiautomatischen Erschließungsverfahrens sollte die Menge der automatisch vorgeschlagenen Deskriptoren begrenzt werden.
- Der **Aufwand beim Aufbau fachteilgebietsspezifischer MindServer-Versionen** stünde **nicht im Verhältnis zu dem zu erwartenden Qualitätsgewinn**.
- Die **Einführung eines cut-off Levels** sollte in einer **Anschlussstudie vor dem Hintergrund der spezifischen Erschließungsrichtlinien** untersucht werden.

Vielen Dank
für Ihre Aufmerksamkeit!

Bibliographie (1/2)

Bertram, Jutta (2005) Einführung in die inhaltliche Erschließung. Grundlagen – Methoden – Instrumente. Würzburg: ERGON-Verlag.

Deutsche Nationalbibliothek (DNB) (2010) PETRUS. Prozessunterstützende Software für die digitale Deutsche Nationalbibliothek. Interne Präsentation, Mai 2010.

Gerards, Michael/Gerards, Andreas/Weiland, Peter (2006) Der Einsatz der automatischen Indexierungssoftware AUTINDEX im Zentrum für Psychologische Information und Dokumentation (ZPID) http://www.zpid.de/download/PSYNDEXmaterial/autindex.pdf Zugriff am 02.04.2011.

Gerards, Michael (2011) Semiautomatische Erschließung von Psychologie-Information. Präsentation im Rahmen des PETRUS-Workshops an der Deutschen Nationalbibliothek Frankfurt/Main, 21./22. März 2011 http://files.d- nb.de/pdf/petrus/semi-automatische_erschliessung_zpid.pdf Zugriff am 22.05.2011.

GESIS – Informationszentrum Sozialwissenschaften (Hg.) (2005) Regelwerk für die Literaturdokumentation Sozialwissenschaften. Regeln für die inhaltliche Erschließung sozialwissenschaftlicher Literatur.

Groß, Thomas (2010) Die Implementierung eines automatischen Indexierungsverfahrens am Beispiel der Deutschen Zentralbibliothek für Wirtschaftswissenschaften. Masterarbeit im Rahmen des postgradualen Fernstudiums Master of Arts. http://edoc.hu-berlin.de/master/gross-thomas-2010- 05-08/PDF/gross.pdf Zugriff am 10.11.2010.

Groß, Thomas/Faden, Manfred (2010) Automatische Indexierung elektronischer Dokumente an der Deutschen Zentralbibliothek für Wirtschaftswissenschaften. In: Bibliotheksdienst, 44. Jg., 12, 1120-1135.

Informationszentrum Sozialwissenschaften (IZ) (2002) Thesaurus Sozialwissenschaften. Alphabetischer Teil/Systematischer Teil. Bonn: Informationszentrum Sozialwissenschaften.

Informationszentrum Sozialwissenschaften (IZ) (2006) Thesaurus Sozialwissenschaften. Alphabetischer Teil/Systematischer Teil. Bonn: Informationszentrum Sozialwissenschaften.

Keil, Stefan/Tiesler, Philipp et al. (2010) Automatische Erschließung für die Datenbank SOLIS. Internes Arbeitspapier von GESIS – Leibniz-Institut für Sozialwissenschaften.

Kempf, Andreas Oskar (2012) Automatische Indexierung in der sozialwissenschaftlichen Fachinformation. Eine Evaluationsstudie zur maschinellen Erschließung für die Datenbank SOLIS. Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft, 329. Berlin: Institut für Bibliotheks- und Informationswissenschaft der Humboldt-Universität zu Berlin.

Bibliographie (2/2)

Krause, Jürgen (1996) Informationserschließung und -bereitstellung zwischen Deregulation, Kommerzialisierung und weltweiter Vernetzung – Schalenmodell. IZ- Arbeitsbericht Nr. 6, Bonn: Informationszentrum Sozialwissenschaften.

Krause, Jürgen (2006) Shell Model, Semantic Web and Web Information Retrieval. In: Harms, Ilse/Luckhardt, Heinz-Dirk/Giessen, Hans W. (Hg.) Information und Sprache. Beiträge zu Informationswissenschaft, Computerlinguistik, Bibliothekswesen und verwandten Fächern. Festschrift für Harald H. Zimmermann. K.G. Saur: München, 95-106.

Nohr, Holger (2004[5]) Theorie des Information Retrieval II: Automatische Indexierung, in: Kuhlen, Rainer et al. (Hg.) Grundlagen der praktischen Information und Dokumentation, München: Saur, 215-225.

Ders. (2005[3]) Grundlagen der automatischen Indexierung – Ein Lehrbuch. Berlin: Logos-Verlag.

Normausschuss Bibliotheks- und Dokumentationswesen (1988) DIN 1426. Inhaltsangaben von Dokumenten; Kurzreferate, Literaturberichte. Berlin u. a.: Beuth.

Organisation der Vereinten Nationen für Erziehung, Wissenschaft und Kultur – Büro für internationale Normen und Rechtsfragen (Hg.) (1979) Empfehlungen zur internationalen Vereinheitlichung der Statistiken über Wissenschaft und Technologie der UNESCO-Generalkonferenz vom 28.11.1978.

Puzicha, Jan (2009) Informationen finden! Intelligente Suchmaschinentechnologie & automatische Kategorisierung. Technical Whitepaper – Grundlagen der Informationsgewinnung. MindServer. Publikation der Firma Recommind.

Schöning-Walter, Christa (2011) Automatische Erschließungsverfahren für Netzpublikationen. Stand der Arbeiten im Projekt PETRUS. Präsentation bei einem Kolloquium von GESIS – Leibniz-Institut für Sozialwissenschaften, Februar 2011.

Siegmüller, Renate (2007) Verfahren der automatischen Indexierung in bibliotheksbezogenen Anwendungen. In: Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft, Heft 214. Berlin. http://www.ib.hu-berlin.de/~kumlau/handreichungen/h214/h214.pdf Zugriff am 03.04.2011.

Stock, Wolfgang G./Stock, Mechtild (2008) Wissensrepräsentation. Informationen auswerten und bereitstellen. München: Oldenbourg Verlag.

Support Vector Machine – Wikipedia http://de.wikipedia.org/wiki/Support_Vec-tor_Machine Zugriff am 09.04.2011.

Wissel, Verena (2011) Erfahrungsbericht und Schlussfolgerungen des DIPF zur Erprobung automatischer Erschließung. Präsentation im Rahmen des PETRUS-Workshops an der Deutschen Nationalbibliothek Frankfurt/Main, 21./22. März 2011. http://files.d-nb.de/pdf/petrus/dipf_erfahrungsbericht.pdf Zugriff am 15.05.2011.

Womser-Hacker, Christa (2004[5]) Theorie des Information Retrieval III: Evaluierung, in: Kuhlen, Rainer/Seeger, Thomas/Strauch, Dieter (Hg.) Grundlagen der praktischen Information und Dokumentation, München: K.G. Saur, 227-235.