

Google Scholar: Warum eine (akademische) Suchmaschine nicht ausreicht

Philipp Mayr

GESIS - Leibniz-Institut für Sozialwissenschaften, Bonn

WissenschaftlerInnen stehen heute eine Vielzahl unterschiedlichster Suchmaschinen für die Suche nach wissenschaftlichen Dokumenten zur Verfügung. Neben den klassischen Informationsanbietern Bibliothek, Fachinformation und Verlag sind Web-Suchmaschinen inzwischen fester Bestandteil bei der Recherche nach frei zugänglichen Dokumenten. Scirus [1] und Google Scholar [2] sind zwei Beispiele für Suchdienste kommerzieller Suchmaschinen-Unternehmen, die eine Einschränkung auf den wissenschaftlichen Dokumentenraum anstreben und nennenswerte Dokumentenzahlen in allen Disziplinen generieren. Der Vergleich der Treffermengen für beliebige Suchthemen zeigt deutlich, dass es mengenmäßig einen großen Unterschied macht, mit welchem Suchsystem, in welchem Dokumentenpool und nach welchen Dokumenttypen gesucht wird. Tabelle 1 verdeutlicht dies am Beispiel der Trefferergebnisse für die Suchbegriffe *search engines* bzw. *Suchmaschinen* in der allgemeinen Internet-suchmaschine Google, der wissenschaftlichen Suchmaschine Google Scholar (GS) und der größten fachübergreifenden bibliographischen Literaturdatenbank Web of Science (WoS) [3]. Der Anteil der Dokumente, die in diesem Fall der Wissenschaft zu zuordnen sind (siehe GS und WoS in Tabelle 1), liegt gegenüber der allgemeinen Websuche lediglich im Promille-Bereich.

	Google	GS	WoS	GS/Google in Promille	WoS/Google in Promille	WoS/GS in Promille
search engines	49,000,000	621,000	2,061	12.7	0.0	3.3
Suchmaschinen	6,580,000	9,550	0	1.5	0.0	0.0

Tabelle 1: Vergleich der Trefferzahlen von Google, Google Scholar (GS) und Web of Science (WoS) (abgefragt am 30.06.2008)

Google Scholar

Der Start der akademischen Suchmaschine Google Scholar hat nach der Veröffentlichung im November 2004 wie üblich ein breites Medienecho nach sich gezogen. Sowohl in der allgemeinen Presse als auch unter Wissenschaftlern, Fachverlagen und Wissenschaftsgesellschaften hat Google Scholar insbesondere wegen der Nähe zu den aktuell viel diskutierten Themen Open Access und Invisible Web für Aufsehen gesorgt. Die Besonderheit von Google Scholar liegt neben der zugrunde liegenden Technologie sicherlich in seiner Bemühung nur wissenschaftliche und qualitätsgeprüfte Dokumente zu durchsuchen. Google Scholar gibt dazu Folgendes auf seinen Seiten an: „Google Scholar provides a simple way to broadly search for scholarly literature. From one place, you can search across many disciplines and sources: peer-reviewed papers, theses, books, abstracts and articles, from academic publishers, professional societies, preprint repositories, universities and other scholarly organizations. Google Scholar helps you identify the most relevant research across the world of scholarly research.” [4].

Google begibt sich mit seiner „Wissenschaftssuchmaschine“ in einen bislang von Bibliotheken und Dokumentationseinrichtungen dominierten Bereich. Google hält - wie bei seinen anderen Services übrigens auch - über die Reichweite, Aktualität und Abdeckung von Google

Scholar bedauerlicherweise keine Informationen bereit. Daher wurde im August 2006 im Rahmen des GESIS-Projekts „Kompetenzzentrum Modellbildung und Heterogenitätsbehandlung“ [5] (zum Projekt siehe Mayr/Petras, 2008) empirisch untersucht, wie vollständig Google Scholar den Bereich der wissenschaftlichen Fachinformation in unterschiedlichen Bereichen erfasst. Der Suchdienst hat sich in der Zwischenzeit durchaus verbessert, die folgenden Grundaussagen sind tendenziell aber immer noch gültig. In der Untersuchung (Mayr/Walter, 2007) wurde der Umfang des Services anhand der Abdeckung fachlich ausgerichteter Zeitschriftenlisten gemessen (siehe dazu Tabelle 2):

- a) drei Zeitschriftenlisten von Thomson Scientific [6]: Arts & Humanities Citation Index (A&HCI), Social Science Citation Index (SSCI), Science Citation Index (SCI),
- b) die Open Access-Zeitschriften des Verzeichnisses DOAJ [7],
- c) in der Fachdatenbank SOLIS ausgewertete sozialwissenschaftliche Zeitschriften (GESIS) [8].

Weiterhin wurde untersucht, welche Typen von Nachweisen und welche Webserver sich in den analysierten Trefferdaten befanden.

Die Ergebnisse der Studie zeigen, dass sich ein Großteil der Zeitschriften der abgefragten Zeitschriftenlisten in Google Scholar identifizieren lassen. Genauer betrachtet, wird dieses Ergebnis jedoch durch den hohen Anteil an nicht weiterführenden Literaturangaben sog. „Citations“ relativiert (der Dokumenttyp Citations macht insgesamt 28% über alle Zeitschriftenlisten aus, siehe dazu auch Tabelle 2). Die internationalen Zeitschriften des SCI von Thomson Scientific aus dem Bereich STM sind vergleichsweise gut abgedeckt (61% weiterführende Links).

Liste	Link %	Citations %	Full-text %
A&HCI	41.78	50.73	7.49
DOAJ	48.29	29.61	22.11
GESIS	10.42	83.11	6.48
SCI	61.35	16.72	21.94
SSCI	49.38	32.84	17.78

Tabelle 2: Verteilung der Dokumenttypen über die untersuchten Zeitschriftenlisten (aus Mayr/Walter, 2007)

Der deutschsprachige Anteil an wissenschaftlichen Zeitschriften in Google Scholar, getestet anhand der sozialwissenschaftlich ausgerichteten Zeitschriftenliste der GESIS (83% Citations), ist aller Wahrscheinlichkeit nach eher gering und unvollständig. Die Analyse der Webserver zeigt, dass vorrangig die Fachangebote von kommerziellen Wissenschaftsverlagen wie z.B. Springer, Ingenta, Wiley usw. (allerdings nicht vollständig) indexiert wurden. Unsere Ergebnisse verdeutlichen, dass umfangreiche elektronisch frei zugängliche Bestände, insbesondere aus dem Open Access (siehe DOAJ-Liste) und self-archiving-Bereich bislang zu wenig berücksichtigt wurden. Unverständlich ist, dass Zeitschriftenartikel, die sich auf frei im Internet verfügbaren Webservern befinden, häufig von Google Scholar nicht nachgewiesen werden, obwohl sie meistens über eine klassische Google-Suche zu finden sind. Google kündigt an, „scholarly articles across the web“ anzubieten, dafür ist der Anteil der nachgewiesenen Artikel aus Open Access-Zeitschriften bzw. der selbstarchivierten Volltexte (Eprints, Preprints) zu gering. Unsere Tests bestätigen weiterhin, dass Google Scholar in vielen Dokumentkollektionen keine tagesaktuellen Daten präsentieren kann und die Trefferdaten aufgrund der Implementation der automatischen Zitationsextraktion (vgl. Lawrence et al, 1999) z. T. unvollständig, fehlerhaft und häufig redundant aufgelistet werden (vgl. Jacsó, 2006). Des Weiteren werden wie auch bei Scirus nichtwissenschaftliche Quellen (z.B. studentische Seminararbeiten usw.) in den Trefferdaten nachgewiesen.

Fazit

Wie der bekannte Suchdienst Google Web Search bietet auch Google Scholar die gewohnt schnelle Suche und eine einfach zu bedienende Benutzeroberfläche. Pluspunkte sind, dass die Recherche kostenfrei ist und dass im Volltext interdisziplinärer Bestände gesucht werden kann. Der Ansatz von Google Scholar bietet für Literatursuchende einige Potenziale, wie z.B. die automatische Zitationsanalyse und das darauf aufbauende Ranking und Browsing sowie in vielen Fällen den direkten Volltextzugriff. Evaluation von Zitationszahlen oder webometrische Untersuchungen (Thelwall et al., 2005) auf Basis der Google Scholar Daten sind aufgrund der kostenfreien Nutzung des Services u. U. fruchtbar, allerdings aufgrund der Vagheit in den Daten mit großer Vorsicht zu betrachten.

Im Vergleich zu Fachdatenbanken mit ihren hohen Anforderungen an die Dokumentenqualität (z.B. nur peer-reviewed papers in WoS) sowie der Fokussierung auf Precision und Recall (s. Mayr/Petras, 2008) bietet Google Scholar momentan nicht die Transparenz und Vollständigkeit, die viele Nutzer von einem wissenschaftlichen Informationsangebot erwarten. Als Ergänzung der Recherche in Fachdatenbanken - v. a. durch die teilweise Abdeckung einer Reihe von Open Access-Zeitschriften - kann Google Scholar aber durchaus nützlich sein.

Literatur

- Jacsó, Péter (2006): Deflated, Inflated and Phantom Citation Counts. In: Online Information Review 30, No. 3, pp. 297-309
- Lawrence, Steve; Giles, C. Lee; Bollacker, Kurt (1999): Digital Libraries and Autonomous Citation Indexing. In: IEEE Computer 32, No. 6, pp. 67-71. URL: <http://citeseer.ist.psu.edu/aci-computer/aci-computer99.html>
- Mayr, Philipp; Petras, Vivien (2008): Building a terminology network for search: the KoMoHe project. In: International Conference on Dublin Core and Metadata Applications. Berlin
- Mayr, Philipp; Petras, Vivien (2008): Cross-concordances: terminology mapping and its effectiveness for information retrieval. In: IFLA World Library and Information Congress. Québec, Canada URL: http://www.ifla.org/IV/ifla74/papers/129-Mayr_Petras-en.pdf
- Mayr, Philipp; Walter, Anne-Kathrin (2007): An exploratory study of Google Scholar. In: Online Information Review 31, No. 6, pp. 814-830. Preprint available at <http://arxiv.org/abs/0707.3575>
- Thelwall, M.; Vaughan, L.; Björneborn, L. (2005): Webometrics. In: Annual Review of Information Science and Technology (ARIST) 39, pp. 81-135.

Internetquellen

- [1] <http://scirus.com>
- [2] <http://scholar.google.com>
- [3] <http://scientific.thomson.com/products/wos/>
- [4] <http://scholar.google.de/intl/en/scholar/about.html>
- [5] <http://www.gesis.org/forschung-lehre/programme-projekte/informationwissenschaften/projektuebersicht/komohe/>
- [6] <http://scientific.thomson.com/mjl/>
- [7] <http://www.doaj.org/>
- [8] <http://www.gesis.org/dienstleistungen/fachinformationen/datenbanken-informationssysteme/literaturdatenbank-solis/>