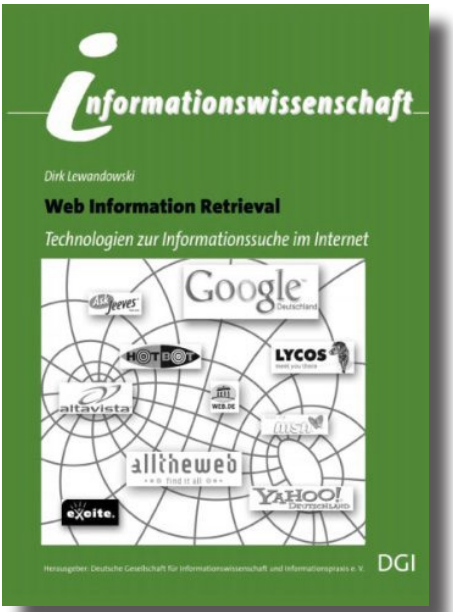


**Dirk Lewandowski: Web Information Retrieval: Technologien zur Informationssuche im Internet. Frankfurt am Main: Deutsche Gesellschaft für Informationswissenschaft und Informationspraxis, 2005. 248 S. (DGI-Schrift; Informationswissenschaft 7).**
**ISBN 3-925474-55-2.**
**(Brosch., EUR 25,00)**

Dies ist die Buchausgabe der am Düsseldorfer Lehrstuhl für Informationswissenschaft angefertigten Dissertation des Autors, der dort auch als Lehrbeauftragter tätig ist und im Sommer 2005 promoviert hat. Trotz seines günstigen Preises ist das Werk auch vollständig in einer (kostenlosen) Online-Ausgabe verfügbar;[1] dennoch hat die DGI aufgrund des bisherigen Interesses an der Papierausgabe bereits einen Nachdruck auflegen können.

Um es gleich vorwegzunehmen: Hier liegt eine erfreulich gut lesbare und sauber gegliederte Bestandsaufnahme des gegenwärtigen Standes der Welt der Web-Suchmaschinen vor. Ziel des Buches ist es, eine Grundlage für das Verständnis der Funktionsweise wie auch der Schwächen dieser Retrievalinstrumente zu bieten – und dies ist dem Autor auch durchaus gelungen. Insbesondere ist er bestrebt, das Information Retrieval (IR) im

Web und dessen Spezifika aus informationswissenschaftlicher Sicht zu diskutieren, vor allem im Hinblick auf einen Vergleich mit dem „klassischen“ IR und den dort eingesetzten Methoden und Modellen. Dass etwa für Retrievaltests im Web gepflegte Testkollektionen mit strukturierten Dokumenten, einer genormten Sacherschließung und a priori erstellten Relevanzurteilen fehlen, liegt ja auf der Hand. Auch mit versierten (professionellen) Benutzern kann hier nicht gerechnet werden. Schwierig ist in diesem Kontext wohl auch, dass die bisherigen Forschungsergebnisse überwiegend proprietär sind und von den betreffenden Eignern (= Anbietern) oft nicht veröffentlicht wurden.

Lewandowskis Buch ist in 14 Kapitel gegliedert, die (abgesehen von Einleitung und Ausblick) zwei Hauptteilen zugeordnet werden können. Der erste bietet eine allgemeine Darstellung des Web-IR in Abhebung vom klassischen IR (Kapitel 2 bis 10), während im zweiten denkbare Möglichkeiten einer schrittweisen Verfeinerung von Recherchen durch Endbenutzer diskutiert werden (Kapitel 11 bis 13).

Der erste Teil beginnt mit einer Abgrenzung des Forschungsumfeldes für das vorliegende Werk. Dabei handelt es sich primär um die Welt der „algorithmischen“ Suchmaschinen wie z.B. Google, deren Aufbau und Abfragesprachen erläutert werden. Der Autor weist darauf hin, dass an Suchmaschinen im Vergleich zum herkömmlichen IR wesentlich heterogenere Abfragen gerichtet werden und referiert die bisher vorliegenden Ergebnisse der Benutzerforschung. Im folgenden Kapitel werden die Größe des Web bzw. die Vollständigkeit seiner Abdeckung durch Suchmaschinen thematisiert; insbesondere wird das so genannte „invisible Web“ näher analysiert. Das mit „Strukturinformationen“ übertitelte vierte Kapitel beschäftigt sich mit dem eher schwachen Strukturierungsgrad von Web-Dokumenten (in den gängigen Formaten wie HTML, Word und PDF) und spricht auch kurz Aspekte wie die Trennung von Navigationselementen, Layout und Inhalt, sowie die Repräsentation der Dokumente in den Datenbanken der Suchmaschinen an.

Kapitel 5 versucht, die Unterschiede zwischen klassischem IR und Web-IR herauszuarbeiten; als wichtigste Unterschiede werden das im Fall des Web fehlende kontrollierte Vokabular sowie die eher bescheidenen Möglichkeiten zur Auswahl von Web-Dokumenten beim Harvesting universeller Suchmaschinen dargestellt. Das Kapitel befasst sich weiter mit den klassischen IR-Modellen (Boolesches, Vektorraum-, probabilistisches Modell), v.a. auch im Hinblick auf ihren Einsatz in Suchmaschinen. Darauf folgt ein kurzes Kapitel über Ranking, das sich mit den bei Suchmaschinen eingesetzten Rankingfaktoren sowie der Problematik der Messbarkeit von

Relevanz befasst. Mit Ranking bzw. Relevanzverbesserung haben auch die im folgenden Kapitel 7 dargestellten Verfahren der Informationsstatistik (wortfrequenzbasierte und nutzungsstatistische Verfahren) und Informationslinguistik (bei Suchmaschinen noch wenig angewandt) zu tun, wie auch die im nächsten Kapitel diskutierten, auf dem klassischen Citation Indexing basierenden linktopologischen Rankingansätze. Hier werden neben Googles *PageRank* auch die Verfahren *HITS* und *Hilltop* vorgestellt und im Hinblick auf ihre Evaluierbarkeit und verschiedene Probleme diskutiert. Im neunten Kapitel geht es um Retrievaltests (klassisches IR; Recall/Precision) und die Resultate ausgewählter Suchmaschinen-Tests, wobei eine nur geringfügige Anpassung der Testmethodik an die Gegebenheiten von Web und Suchmaschinen konstatiert wird. Da neben der ermittelten Precision auch andere Faktoren die Qualität von Suchmaschinen bestimmen, wird im letzten Kapitel des ersten Teils auf Verfahren der intuitiven Benutzerführung eingegangen (relevance feedback, query expansion, Einsatz von Klassifikation und Thesaurus, Clustering und graphische Veranschaulichung von Ergebnissen) und auf Beispiele aus dem Suchmaschinensektor Bezug genommen.

Im zweiten, kürzeren Teil des Buches werden Verbesserungsmöglichkeiten thematisiert, die davon ausgehen, dass die Nutzer von Web-Suchmaschinen in größerem Ausmaß Kontrolle über die Treffermengen erhalten sollen, wobei sowohl Laien als auch Profi-Rechercheure in die Betrachtung eingeschlossen werden. Als Kernbereiche für solche Verbesserungen identifiziert der Autor die Aspekte Aktualität, Qualität und Dokumentrepräsentation. Dazu werden in Kapitel 11 die spezifischen Probleme der Datumsbeschränkung in Suchmaschinen ins Visier genommen – ein bislang nicht wirklich gelöster Problembereich (scheinbare vs. tatsächliche Aktualisierung von Dokumenten, z.B. im Hinblick auf die Auswirkungen beim Ranking); eine Datumsbeschränkung könnte künftig auch durch Verfahren der intuitiven Benutzerführung in die Trefferlisten eingebaut werden. Das folgende Kapitel 12 beschäftigt sich mit den Möglichkeiten der Ermittlung besonders hochwertiger Quellen aus dem Web, wobei allerdings grundsätzlich eine menschliche Intervention vorausgesetzt wird (bewertende Auszeichnung der Dokumente). Lewandowski führt in diesem Zusammenhang die aus dem Bereich großer Online-Hosts bekannten „Top-Quellen" in die Betrachtung ein und bringt Beispiele für die Einbindung ähnlicher autoritativer Ressourcen in Suchmaschinen bzw. deren Ergebnisdarstellungen. Auch die Einbindung von Quellen des als qualitativ hochwertig angenommenen „invisible Web" und aus Web-Katalogen werden hier angesprochen.

Das dritte Kapitel des zweiten Teils hat die Verbesserung der Repräsentation der von den Suchmaschinen indexierten Dokumente durch weitere, aus dem Inhalt der letzteren gewonnene Attribute zum Thema. Dazu zählen bspw. die Entfernung der nicht zum Inhaltsteil der Dokumente gehörenden Teile (z.B. Navigationselemente oder Werbung), die Identifizierung des tatsächlichen Titels des Dokuments (anstelle der Verwendung bloß des <title>-Tags) sowie realistischer (d.h. um den obigen Ballast bereinigter) Größenangaben. Auch die Problematik der Verwendung von Web-Domänen für die Einschränkung auf bestimmte Arten von Quellen (z.B. Dokumente aus dem akademischen Bereich) fällt in diesen Bereich, wie auch die Art der Komprimierung der Dokumente zu den in Trefferlisten dargestellten Kurzversionen. Das Buch schließt sodann mit einem kurzen Fazit („... Web Information Retrieval zu einem beträchtlichen Teil nur wenig erforscht ...“), einem Literaturverzeichnis und einem Sachregister.

Dass Lewandowskis Überblicksdarstellung im Prinzip gut gelungen und lesbar ist, wurde schon eingangs festgehalten. Hier sei aber doch noch ein kritischer Kommentar erlaubt – das Buch ist in etlichen Abschnitten ganz einfach viel zu knapp! Oft hatte ich den Eindruck, einen geradezu um besondere Ökonomie der Darstellung bemühten Text zu lesen, vielfach hätte ich mir mehr Details und eine intensivere Auseinandersetzung gewünscht. Dies bezieht sich nicht auf die (allerdings fast unumgängliche) Darstellung von Lehrbuchwissen wie etwa der IR-Modelle oder der Verfahren der Informationslinguistik, sondern auf alles, was mit der „Schnittstelle“ zwischen klassischem IR und der Welt der Suchmaschinen zu tun hat. Zudem hätte ich mir die im zweiten Teil diskutierten Verbesserungsmöglichkeiten im Rahmen einer Dissertation vielleicht auch in der einen oder anderen Form *realisiert* bzw. *empirisch überprüft* vorstellen können. Diese Einwände sollen aber einer grundsätzlichen Empfehlung des Buches keinen Abbruch tun. Ich rate allen im Bibliotheks- und Informationsbereich Tätigen, die Interesse am Web und seinem Informationspotential haben, es selbst zu lesen und sich mit den darin angesprochenen Inhalten intensiv auseinanderzusetzen!

Otto Oberhauser, Wien

[1] http://www.durchdenken.de/lewandowski/web-ir/