
ex.t Karrieretipps
Mechtild und Wolfgang G. Stock
Recherchieren im Internet
expert verlag

Bibliografische Information Der Deutschen Bibliothek

Die Deutsche Bibliothek verzeichnet diese Publikation
in der Deutschen Nationalbibliografie;
detaillierte bibliografische Daten sind im Internet über
http://dnb.ddb.de abrufbar.

Bibliographic Information published by Die Deutsche Bibliothek

Die Deutsche Bibliothek lists this Publication
in the Deutsche Nationalbibliografie;
detailed bibliographic data is available in the Internet at
http://dnb.ddb.de .

ISBN 3-8169-2278-3

Bei der Erstellung des Buches wurde mit großer Sorgfalt vorgegangen; trotzdem können Fehler nicht vollständig ausgeschlossen werden. Verlag und Autoren können für fehlerhafte Angaben und deren Folgen weder eine juristische Verantwortung noch irgendeine Haftung übernehmen.
Für Verbesserungsvorschläge und Hinweise auf Fehler sind Verlag und Autoren dankbar.

© 2004 by expert verlag, Wankelstr. 13, D-71272 Renningen
Tel.: +49 (0) 7159 - 9265-0, Fax +49 (0) 7159 - 9265-20
E-Mail: expert@expertverlag.de, Internet: www.expertverlag.de
Printed in Germany

Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung außerhalb der engen Grenzen des Urheberrechtsgesetzes ist ohne Zustimmung des Verlags unzulässig und strafbar. Dies gilt insbesondere für Vervielfältigungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Verarbeitung in elektronischen Systemen.

Inhaltsverzeichnis

Vorwort

Bücher der Art „Wie bediene ich Suchmaschinen im Web?“ gibt es genügend. Nach Art guter Kochbücher erläutern sie, was zu tun ist, will man erfolgreich nach Informationen suchen. Einen theoretischen Hintergrund geben solche Bücher nicht; dies ist auch sehr legitim, hier kommt es ausschließlich auf die Praxis an. Artikel in Fachzeitschriften und Konferenzproceedings sowie Patente zum Information Retrieval existieren zu Tausenden. Theoretische wie praktische Details der informatischen, computerlinguistischen oder informationswissenschaftlichen Grundlagen und der technischen Algorithmen werden minutiös beschrieben und erklärt. Einen für den Laien lesbaren Stoff gibt diese Literatur meist nicht ab; ebenso wird nicht immer der Praxisbezug deutlich, ja ist manchmal gar nicht angestrebt. Auch diese theoretische Literatur ist sehr wichtig, treibt sie doch die Entwicklung im Information Retrieval voran. Die Situation ist für den „Informationslaien“ nicht befriedigend: Einmal erhält er eine Praxisdiät, zum andern eine Theoriediät. Der Internetnutzer, der – zumindest punktuell – wissen möchte, wie ein Suchwerkzeug arbeitet und welcher Informationsinhalt dort gespeichert ist, wird kaum geeignete Werke finden. In diese Lücke stößt unser kleines Büchlein. Es stellt anhand von Fallbeispielen dar, was „hinter den Kulissen läuft“, es bespricht verständlich (hoffentlich!) einzelne Recherchewerkzeuge im Internet, dies jedoch stets kritisch in der Perspektive der Informationswissenschaft und –praxis.

Viele Internetnutzer verlassen sich bei der Informationssuche auf die „großen“ Suchwerkzeuge wie Yahoo! oder Google. Wir wollen dem Leser zeigen, dass dies *erstens* ein guter Einstieg ist, aber *zweitens* in den meisten Fällen absolut nicht ausreicht.

Wir beginnen unseren Streifzug mit Suchwerkzeugen im World Wide Web. Zunächst soll der Unterschied zwischen einem Webkatalog (wie Yahoo!) und einer Suchmaschine deutlich werden. Bei den Suchmaschinen zeigen wir, dass ein entscheidendes Moment die Rangordnung der Treffer, das sog. „Relevance Ranking“ ist. Nach der Art, wie solche Rankings zustandekommen, unterscheiden wir zwischen sprachwissenschaftlichen Verfahren (wie bei AltaVista) und Verfahren, die die Web-Topologie, also die Linkstruktur im WWW ausnutzen (wie Google). Am Rande gehen wir darauf ein, wie Rangplätze gekauft bzw. – dies ist häufigste Form – ersteigert werden.

Eine Spezialsuchmaschine von Google ist Froogle, eine Suchmaschine für Produkte. Im Kontrast mit einer professionellen Produktdatenbank (Wer liefert was?) lassen sich sehr schön die Vorteile eigener Datenbanken ausmachen. Solche Informationen erhält

man nicht direkt bei den Suchmaschinen, sondern bei einer Spezialdatenbank im sog. „Deep Web". Ist Wer liefert was? für den Webnutzer kostenlos, so stoßen wir in den tieferen Gefilden des Deep Web häufig auf Kassenhäuschen, sprich hier wird für die Informationen gezahlt – und dies zu recht, da hohe Kosten beim Sammeln sowie Auswerten der Informationen entstehen und damit diese Dienste erst möglich werden.

Als Unternehmen steht man häufig vor dem Problem, eine neue Geschäftsverbindung eingehen zu müssen. Aber ist der neue Kunde oder der neue Zulieferer auch ein sicherer Partner? Wie ist es um seine Bonität bestellt? Im Deep Web finden sich gesicherte Bonitätsinformationen. Wir zeigen, wie der deutsche Marktführer in diesem Segment des Informationsmarktes – Creditreform – seine Firmendossiers erstellt und wie das entsprechende Retrievalsystem aussieht.

Wer liefert was? und Creditreform sind singuläre Datenbanken im Web. Es gibt Unternehmen, die hunderte bis tausende singuläre Datenbanken bündeln und zu einem System aggregieren. In einem solchen One-Stop-Shop erhält der Kunde unter einer Oberfläche Zugang zu diversen qualitativ gesicherten Informationen. Hier haben wir den deutschen Anbieter von Wirtschaftsinformationen, die GBI, als Beispiel ausgesucht.

Was ist die hauptsächliche Botschaft dieses Buches? Das World Wide Web ist für die Suche nach Informationen eine Goldgrube, allerdings nur, wenn man geschickt die Angebote der Suchwerkzeuge im Oberflächenweb und die Datenbanken sowie die Content-Aggregatoren im Deep Web kombiniert durchsucht. Machen wir uns nichts vor: Die meisten Informationen, die Google und Co. finden, haben das Flair von Hochglanzkatalogen (im günstigen Fall). Erst mit den Suchwerkzeugen im Deep Web und dessen Qualitätsinformationen wird eine Recherche im Internet zufriedenstellend.

Im World Wide Web ist Bewegung. Man kann sich nicht darauf verlassen, dass eine Website, eine Suchmaschine oder auch ein Informationsangebot im Deep Web morgen noch so aussieht wie heute. Insofern bitten wir unsere Leser um Nachsicht, falls die von uns beschriebenen Systeme zum Recherchieren im Internet inzwischen geändert worden sind. Stand der Dinge ist der März 2003.

Die Kapitel 1 bis 4 sowie 7 sind in einer anderen Form bereits im deutschen Newsletter zur Informationswirtschaft „Password" erschienen (natürlich haben wir sie aktualisiert). Wir danken dem Chefredakteur von „Password", Dr. Willi Bredemeier, dass wir die Texte an dieser Stelle neu drucken lassen dürfen. Weiterhin danken wir Creditreform und der GBI für die bereitwillige Überlassung kostenloser Passworte für die Analyse und Bewertung ihrer Informationsprodukte.

Mechtild Stock (MechtildStock@aol.com)
Wolfgang G. Stock (Stock@phil-fak.uni-duesseldorf.de)

1 Recherchieren im Internet: An der Oberfläche fischen oder im Deep Web tauchen?

Warum dieses Buch?

Das Internet beinhaltet weitaus mehr Informationen als gemeinhin angenommen. Neben den Informationen im „Oberflächenweb", die von Suchwerkzeugen zumindest partiell gesammelt werden, existieren im „Deep Web" sozusagen im Verborgenen sowohl kostenlos abfragbare Datenbanken als auch die digitalen Informationsdienste der (kostenpflichtigen) professionellen Informationsanbieter. Nur eine Recherche in der Gesamtheit der Informationen *im* Web und derjenigen, die zusätzlich *über* das Web erreicht werden, ist zielführend und hilft, erkannte Wissenslücken zu schließen, unternehmerische Entscheidungen vorzubereiten sowie Frühwarnsignale zu erhalten.

Das Buch unternimmt anhand von Fallbeispielen einen Streifzug durch die unterschiedlichen Weltregionen des Internet. Wir fragen jeweils danach, wie die Suchwerkzeuge und Datenbanken arbeiten, wie sie zu ihren Dokumenten kommen, wie diese inhaltlich ausgewertet werden, wie die Rangfolgen in den Ergebnislisten zustande kommen und wie die Benutzerschnittstellen aussehen. Dieses Hintergrundwissen ist notwendig, damit wir – als Nutzer – mit den Suchwerkzeugen und Datenbanken optimal arbeiten lernen. Leider (oder zum Glück – je nach Sichtweise) gehört zu einer zufriedenstellenden Suche im Internet mehr dazu, als einige Worte in ein Suchfenster einer Suchmaschine einzutragen.

Die Suchwerkzeuge des Oberflächenweb zerfallen in Webkataloge und in Suchmaschinen. Bei letzteren unterscheiden wir weiterhin nach sprachwissenschaftlich fundierten Systemen und nach Systemen, die nach der Popularität von Webseiten gewichten. Unser Fallbeispiel eines Webkatalogs ist Yahoo!, als linguistisch orientierte Suchmaschine beschreiben wir AltaVista, als Suchmaschine auf der Basis von Popularität analysieren wir Google. Da das Deep Web ausgesprochen umfangreich und unübersichtlich ist, konzentrieren wir uns auf deutschsprachige Angebote. Unser Fallbeispiel einer kostenlosen Datenbank ist „Wer liefert was?", ein digitaler Katalog von B-to-B-Produkten; sodann wenden wir uns mit Creditreform einem kostenpflichtigen Angebot zu. Das Beispiel eines professionellen Content-Aggregators gibt GBI – the contentmachine.

Zielgruppe des Buches sind Mitarbeiter von Unternehmen und anderen Institutionen, die beruflich im Internet recherchieren oder dies in Zukunft tun wollen und denen ihre Arbeitszeit zu knapp ist, um umständlich mit Suchwerkzeugen zu experimentieren und danach trotzdem nur ein suboptimales Rechercheergebnis vorzulegen. Das Buch ist so

geschrieben, dass der Leser in recht kurzer Zeit einen Einblick in die Arbeitsweisen der Systeme im Web findet und – dank der detaillierten Fallbeispiele - dieses Wissen auch in der Praxis umsetzen kann.

„Oberfläche" und „Tiefenregionen" des Internet

Mit dem Buch „The Invisible Web" von Chris Sherman und Gary Price wurde für viele Laien klar, was für Information Professionals wohl kaum außer Frage stand: Suchwerkzeuge im Internet – Suchmaschinen wie Google oder Alltheweb sowie Webkataloge wie Yahoo! und das Open Directory – finden nicht alle digitalen Informationen, die weltweit im und über das Internet eigentlich erreichbar wären. Diesen Teil des Internet bezeichnen Sherman und Price als das unsichtbare Web. „In a nutshell, the Invisible Web consists of material that general-purpose search engines either can not, or perphaps more importantly, *will not* include in their collections of Web pages", so Sherman und Price 2001. Da unsere Autoren in ihrem Buch nun aber ausführliche Listen von Einstiegsseiten in das "Invisible Web" vorlegen, kann dieses wohl kaum so unsichtbar sein, wie der Name suggeriert. Insofern ist die Bezeichnung etwas irreführend, so dass die Rede vom „Deep Web", den Tiefenregionen des Internet wahrscheinlich treffender ist. Den Begriff des „Deep Web" verwendet das Unternehmen BrightPlanet, das eine Retrievalsoftware zum „Abfischen" von Oberflächen- und von Tiefenregionen des World Wide Web offeriert.

Das **Oberflächenweb** wird mehr oder minder komplett von den Suchwerkzeugen, den automatisch indexierenden Suchmaschinen und den intellektuell auswertenden Webkatalogen erfasst. Untersuchungen (z.B. Stock/Stock 2000) legen nahe, dass zumindest für gängige Datenformate beim parallelen Ansprechen mehrerer der großen Web-Retrievalsysteme (etwa Google, Alltheweb und AltaVista) weite Teile (bis zu 90%) des World Wide Web findbar sind. (Zumindest, wenn man großzügig unterstellt, dass ein „Treffer" auch noch zählt, wenn man ein bis zwei Clicks nach dem angezeigten Treffer fündig wird.) Vollständigkeit zu erwarten, wäre allerdings maßlos übertrieben. Mit der Geschwindigkeit, in der neue Seiten generiert und alte modifiziert werden, können die Suchwerkzeuge nur moderat mithalten.

Teilweise ignorieren die Suchwerkzeuge aber auch „bewusst" Material, und das findet ausdrücklich Zustimmung. Hier geht es um **Spam**, Informationsmüll, den keiner haben möchte und den auch keiner braucht – einige wenige Perverse und Waffennarren vielleicht ausgenommen. (Dass an dieser Stelle latent die Gefahr einer Zensur lauert, sei zugegeben.)

Die **Tiefenregionen** des Web erschliessen sich dem Nutzer durch das Ansprechen gewisser Datenbanken. Da hier die Webseiten dynamisch erzeugt werden, können diese von den Suchwerkzeugen nicht automatisch aufgefunden werden. Der Nutzer stellt eine Anfrage und bekommt im System seine Antwort in den eigens dafür aufgebauten Seiten (deshalb „dynamisch kreierte" Seiten). In den Datenbasen der jeweiligen

Systeme liegen natürlich die Datensätze vor, so dass sie prinzipiell auch in Suchmaschinen einzupflegen wären. Hierzu bedürfte es Abmachungen zwischen den beteiligten Unternehmen – und die sind (von einigen unten aufgeführten Ausnahmen abgesehen) rar.

Michael K. Bergman betont: „Traditional search engines create their indices by spidering or crawling surface Web pages. To be discovered, the page must be static and linked to other pages. Traditional search engines cannot 'see' or retrieve content in the deep Web – those pages do not exist until they are created dynamically as the result of a specific search. Because traditional search engine crawlers can not probe beneath the surface, the deep Web has heretofore been hidden" (Bergman 2001).

Zur Veranschaulichung des Deep Web nun einige Beispiele! Über die Website des HWWA sprechen wir die Wirtschaftsdatenbank des Hamburgischen Welt-Wirtschafts-Archivs an, über die Site des Deutschen Patent- und Markenamtes die Datenbank DepatisNet oder – unser Beispiel in Kapitel 5 – über WLWonline den umfangreichen B-to-B-Produktkatalog von „Wer liefert was?", wobei die Inhalte weder der HWWA-, der DepatisNet- oder der WLW-Datenbank bei Google und Co. recherchierbar sind. Ähnlich verhält es sich in der kommerziellen Informationswirtschaft, nur hier mit der Zwischenstation bei einem Kassenhäuschen. In diesem Segment finden wir die Anbieter von kostenpflichtigen Qualitätsinformationen, also die Selbstvermarkter (wie z.B. Creditreform Online oder Hoppenstedt Firmendatenbank) und die Content-Aggregatoren (u.a. die sog. „Online-Hosts") vor. Die Informationen dieses Teils des Deep Web haben eines gemeinsam: Sie liegen überhaupt nicht **im** Web, sondern **in** einer Datenbank, die **via** Web erreichbar ist.

Der Umfang des Deep Web ist gemäß Michael K. Bergman rund 400- bis 500mal größer als das Oberflächenweb. Gerechnet wird dabei nicht in der Anzahl von Dokumenten, sondern im Umfang von Speicherplatz. Da Bergman u.a. Datensammlungen wie die des „National Climatic Data Center" (mit rund 370.000 GByte) oder die der NASA (mit 220.000 GByte) mitrechnet, ist die große Zahl nicht mehr ganz so verwunderlich. Weil im Deep Web auch die Kataloge aller Online-Bibliothekskataloge vorkommen, die sich ja durchaus überschneiden, müsste die geschätzte Informationsmenge um die Dubletten bereinigt werden. Da allein ein Informationsanbieter wie Lexis-Nexis mit dem Suchmaschinen-Primus Google nach der Anzahl der Datensätze in etwa gleich zieht, ist es ganz sicher richtig, dass das Deep Web weitaus größer ist als das Oberflächenweb. Orientiert man sich an der Zahl der Datensätze, so ist Bergmans Größenfaktor wohl um eine Zehnerpotenz überschätzt.

Eine Taxonomie der digitalen Online-Information

Das Gesamt der online erreichbaren digitalen Informationen ist zunächst in zwei Hauptklassen einzuteilen:

- Informationen, die im Web liegen (d.h. fest verlinkt sind, wobei keinerlei

Kosten für den Nutzer entstehen) – das **Oberflächenweb**

- Informationen, die in Informationssammlungen (i.d.R. Datenbanken) liegen, wobei die Einstiegsseite der Informationssammlung via World Wide Web erreichbar ist – das **Deep Web**.

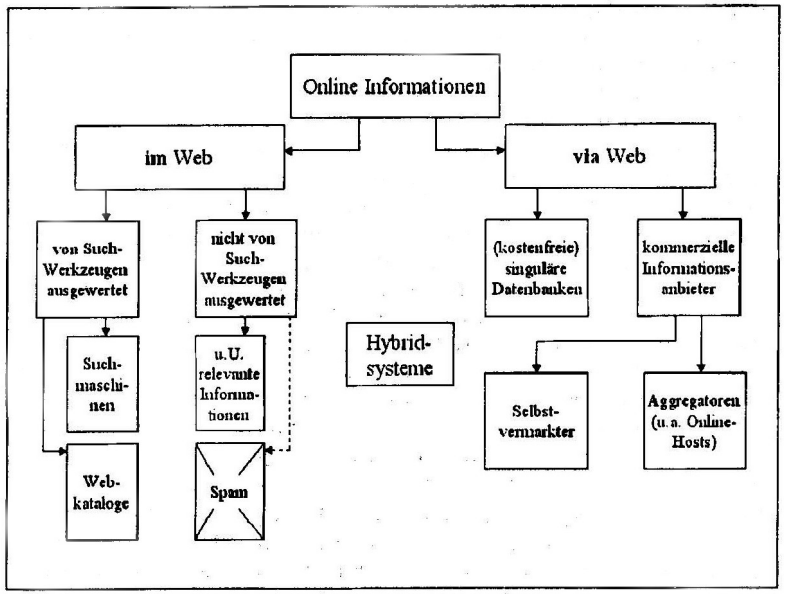

Abbildung 1-1: Welten der Online Informationen im World Wide Web

Die Mehrzahl der Informationen im Web wird von Suchwerkzeugen – nicht von einem einzelnen, aber in der Summe der Werkzeuge – ausgewertet. Nicht immer erfasst werden Informationen mit speziellen Datenformaten, Dokumente, auf die kein Link gerichtet ist; Probleme können u.U. bei der Verwendung von Frames auftauchen – hier droht in der Tat Informationsverlust. (Weiterer Informationsverlust entsteht bei suboptimalem Relevance Ranking von Suchmaschinen: Was nutzt ein Treffer auf Platz 543, den der Nutzer nie sehen wird?) Gewünschter Informationsverlust liegt beim Ausfiltern von Spam vor.

Die zweite Hauptklasse digitaler Online-Informationen liegt nicht selbst im Web, sondern ist nur über das Web erreichbar. Natürlich lassen sich die Einstiegsseiten solcher Systeme auch bei den Suchmaschinen und Webkatalogen finden, nicht aber die

gespeicherten Datensätze. Eine über Suchwerkzeuge vermittelte Suche verläuft indirekt. Zunächst suche ich die Einstiegsseite der Informationssammlung, um im zweiten Schritt beim betreffenden Retrievalsystem mein konkretes Suchargument zu formulieren. Diese via Web erreichbaren Informationen zerfallen wiederum in zwei Unterklassen, und zwar nach der Kostenpflichtigkeit der Angebote. Es gibt singuläre kostenlose Datenbanken (Beispiele: Amazon, HWWA, Medline, diverse Datenbanken nationaler Patentämter, Bibliothekskataloge) und es gibt kostenpflichtige Angebote. Letztere haben als Spezies die Datenbanken der Selbstvermarkter, die genau so singulär dastehen wie die alleinstehenden kostenlosen Datenbanken (nur dass hier halt zu zahlen ist) und als weitere Spezies die Aggregatoren, die verschiedene Informationssammlungen unter einer Oberfläche vereinigen. Aggregatoren können sowohl bei Märkten nicht-digitaler Güter (Beispiele: diverse Shopping Malls), bei Verlagsprodukten (Beispiele: Xipolis als aggregiertes Angebot des Brockhaus und weiterer Nachschlagewerke, Springer Link oder ScienceDirect als Volltextangebote von Zeitschriftenartikeln der beiden Wissenschaftsverlage) als auch in der Form von Online-Hosts auftauchen.

Grenzüberschreitungen

Michael K. Bergman fordert völlig zu Recht: "Clearly, simultaneous searching of multiple surface and deep Web sources is necessary when comprehensive information retrieval is needed" (Bergman 2001). Ob dies nun stets innerhalb genau einer Retrievaloberfläche für alle Informationen vonstatten gehen sollte, kann bezweifelt werden, verliert der Suchende doch dabei die Vorzüge und Extras der einzelnen Datenbanken – das bekannte Problem wie bei den Meta-Suchmaschinen. Meta-Suchmaschinen haben nämlich keine eigenen Datenbasen, sondern ausschließlich ein Retrievalsystem, mit dem sie bei anderen Systemen „schmarotzen" gehen. Da jedes System seine Spezifika hat, haben die Meta-Suchmaschinen (eigentlich) die Aufgabe, die Nutzereingaben in die Syntax der angesprochenen Systeme zu übersetzen. Dies gelingt jedoch nur zu einem kleinen Teil, so dass in der Regel nur der kleinste gemeinsame Nenner der Suchsyntax verwendet wird.

Eine erste Alternative wäre, aus einem einheitlichen System heraus die einzelnen Internetwelten bzw. Datenbanken einzeln anzusteuern und die Dokumente in einen einheitlichen Warenkorb abzulegen. Bei dieser Lösung übernimmt der Nutzer die Hauptarbeit, und er muss die angesprochenen Systeme auch gut kennen. Einzig der Warenkorb hält die unterschiedlichen Aktionen zusammen.

Eine zweite Alternative wird darin bestehen, fachlich begrenzte Ausschnitte zu wählen, über alle Dokumente aller infragekommenden Welten ein kontrolliertes Vokabular zu legen und auch alle Welten mittels dieses Vokabulars abzusuchen. Dies hat den Vorteil der Einheitlichkeit und den möglichen Nachteil der fachlichen Beschränkung.

Wie auch immer die Lösung aussehen wird: Stets ist ein Suchen in mehreren

unterschiedlichen Internetwelten nötig, also ein „Querweltein-Retrieval".

Wenn wir quer durch die Internetwelten Recherchen anbieten, wollen wir von „Hybridsystemen" reden. Ansätze, Hybridsysteme zu kreieren, gibt es bereits vielfach. Je nach der Herkunft des Unternehmens unterscheiden wir folgende Fälle:

- „reiner" Hybrid (Beispiel: Northern Light)
- Suchwerkzeuge im Web mit Querweltein-Ergänzungen (AltaVista.de, Google.com)
- singuläre Datenbank mit Querweltein-Ergänzungen (HWWA)
- kommerzielle Informationsanbieter mit Querweltein-Ergänzungen (Factiva).

Ein vom Ansatz her eindeutiger Hybrid ist **Northern Light**. Dokumente im Web werden konsequent mit proprietärem (d.h. kostenpflichtigem, fachlich einschlägigem) Inhalt innerhalb eines Systems verknüpft. Northern Light ist demnach Suchmaschine und kommerzieller Content-Aggregator in einem System. Abbildung 2 zeigt die Suchoberfläche von NLResearch. Frei zugängliche Webdokumente und Texte des kostenpflichtigen „Special Collection & Premium Content" werden innerhalb einer Datenbank angeboten, zusätzlich gibt es Links zu weiteren einschlägigen Informationssammlungen.

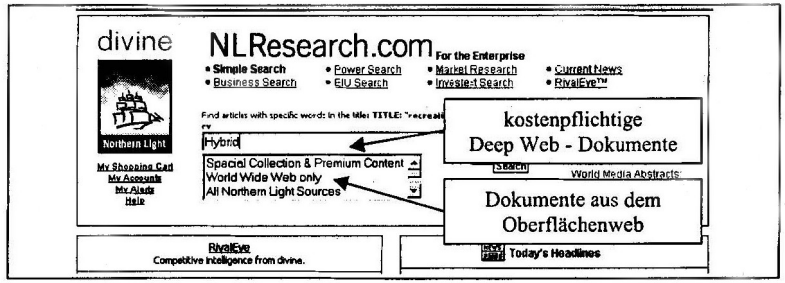

Abbildung 1-2: Hybridsystem Northern Light

AltaVista liefert uns ein Beispiel für eine Web-Suchmaschine, die gewisse Datenbanken in die Suche mit einbezieht. Abbildung 1-3 zeigt eine Trefferliste beim deutschen AltaVista, in der wir (vor die AltaVista-Inhalte sortiert) den Link zur passend zur Suchfrage generierten Trefferliste von Wer liefert Was? finden. Unten auf der Seite liegen Links u.a. zu eBay und Amazon. Unsere Suche nach „Informationsmanagement" bringt demnach über 50.000 Treffer aus der AltaVista-Datenbank, beginnend mit dem Link zur Seite vom Joanneum Graz. Verfolgt man den Link bei Wer liefert was?, so wird man direkt auf die Anzeige der Treffer- oder Suchwortliste bei dieser Datenbank geführt. Der Link zu WLW wird nur dann gezeigt, wenn dort auch relevante Treffer zu

erwarten sind. Die Links etwa zu eBay oder zu Amazon sind immer gelegt (die Suche wird auch automatisch ausgeführt); ein sinnvolles Ergebnis muss aber nicht vorhanden sein. Diese letztere Variante ist aus Nutzersicht nicht sonderlich reizvoll, aber immerhin ein manchmal zielführender Hinweis; für AltaVista ist sie eine Einnahmequelle.

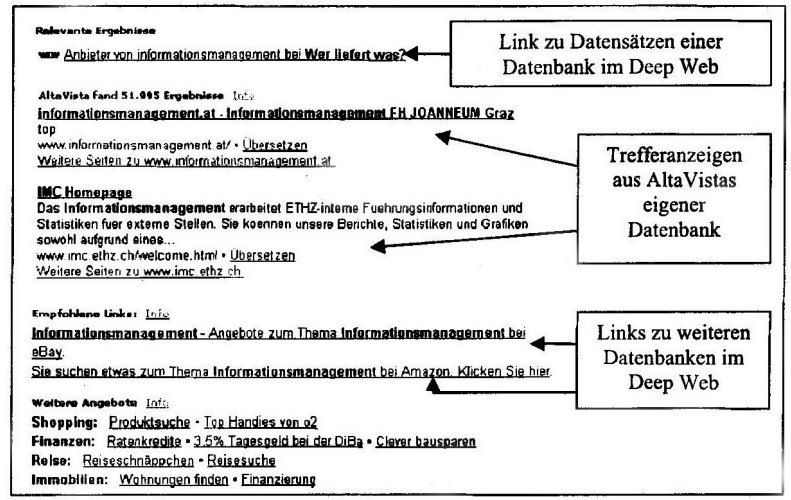

Abbildung 1-3: Suchmaschine AltaVista (deutsch) mit Querweltein-Ergänzungen: Links zu den Datensatz-Treffern singulärer Datenbanken wie Wer liefert Was? nebst Links zu eBay oder Amazon

Ähnlich wie AltaVista geht die us-amerikanische Version von **Google** vor (siehe Abbildung 1-4). Verfolgt man den Link zu Dictionary.com (im Google.com-Anzeigebildschirm sehr versteckt bei der Angabe des unterstrichenen Sucharguments links oben), kommt man zu Lexikoneinträgen sowie zu weiteren Datenbanken, darunter via eLibrary auch zu kostenpflichtigen Volltexten dieses Aggregators. Bei eLibrary liegen Artikel diverser Zeitschriften und Zeitungen in digitaler Form vor. Dies ist eines der wenigen Beispiele bei Suchmaschinen im World Wide Web, in denen auch kostenpflichtiger Informationsinhalt integriert ist.

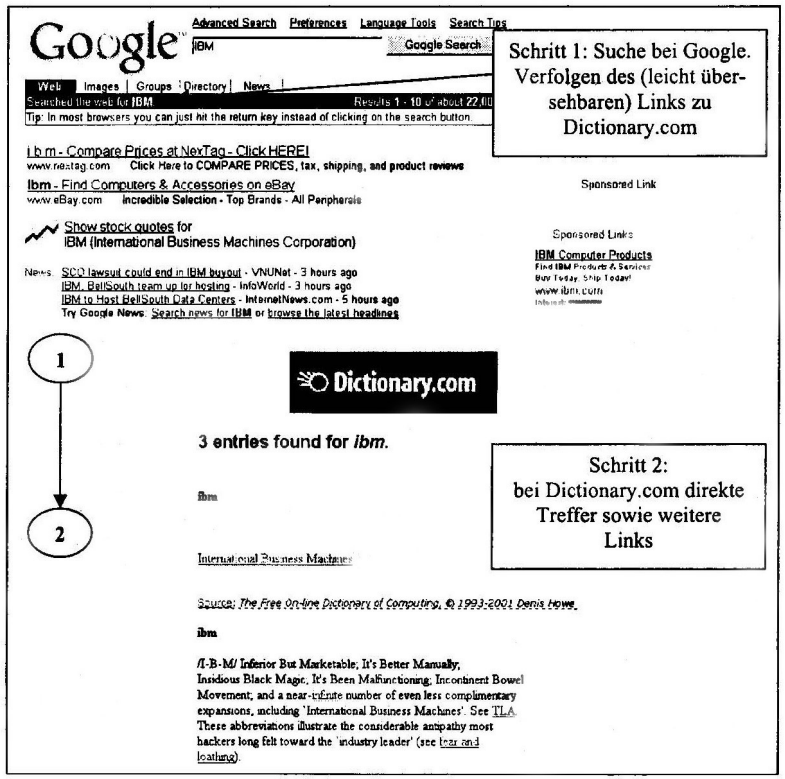

Abbildung 1-4: Google.com mit Querweltein-Ergänzungen zu Dictionary.com und von dort weiter zu diversen – auch kostenpflichtigen – Angeboten

Das **HWWA** versorgt uns mit einem Beispiel der Verknüpfung einer singulären Deep-Web-Datenbank mit dem Oberflächenweb. Die HWWA-Datenbank ist eine Spezialsammlung zur Fachliteratur der Wirtschaftswissenschaft und der Wirtschaftspraxis. Fachspezifische digitale Dokumente werden analog zu gedruckten Dokumenten formal und inhaltlich ausgewertet und als Katalogkarte gespeichert. Die Links der HWWA-Datenbank verweisen sowohl auf proprietäres Material (zu PDF-Volltexten diverser Zeitschriften) als auch (unsere Abbildung 1-5) auf Dokumente im freien Oberflächenweb.

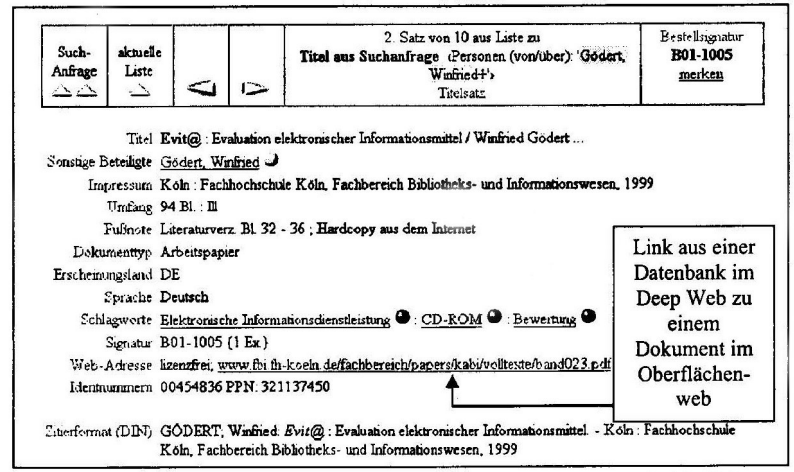

Abbildung 1-5: Datensatzbeispiel der HWWA-Wirtschaftsdatenbank. Verlinkung eines Katalogisats innerhalb einer singulären Datenbank ins Oberflächenweb

Unser letztes Beispiel führt in die kommerzielle Informationswirtschaft. **Factiva**, der Weltmarktführer für News-Informationen (ein Unternehmen von Dow Jones und Reuters) bietet nicht nur proprietären Content an, sondern zusätzlich Dokumente, die im Oberflächenweb liegen (hier meist Newsservices). In Abbildung 1-6 sehen wir eine Trefferliste, die im unteren Teil kostenpflichtige News zeigt, aber zusätzlich (oben) weitere Dokumenttypen, darunter auch (kostenlose) Webseiten, alles natürlich genau passend zur Anfrage des Suchenden.

Erforderlich: Querweltein-Suchen

Querweltein-Retrieval ist nicht nur nötig angesichts der unterschiedlichen Internetwelten, es gibt bereits erste Ansätze, die vormals getrennten Weltregionen der digitalen Informationen gemeinsam zu durchsuchen. Die Bestrebungen der singulären Datenbanken sowie der kommerziellen Informationsanbieter verfolgen dabei einen fachspezifischen Zugang. Dieser hat den Vorteil einer recht zielgenauen Suche, können doch dokumentarische Werkzeuge wie Thesauri bzw. Klassifikationssysteme eingesetzt werden. Die Bestrebungen der Suchmaschinen weiten den Horizont, indem sie die an sie gerichteten Suchargumente an weitere (kostenpflichtige wie kostenlose) singuläre Datenbanken weiterreichen. Hier liegt der Vorteil in der Breite der Datenbasis, wobei die Qualität der zielgenauen Suche leidet. Strategie der Suchmaschinenfirmen sollte sein, die Anzahl der durchzuschaltenden singulären Datenbanken zu maximieren,

zumindest aber solche Datenbanken zu erreichen, bei denen ein „allgemeines Interesse" vorausgesetzt werden kann.

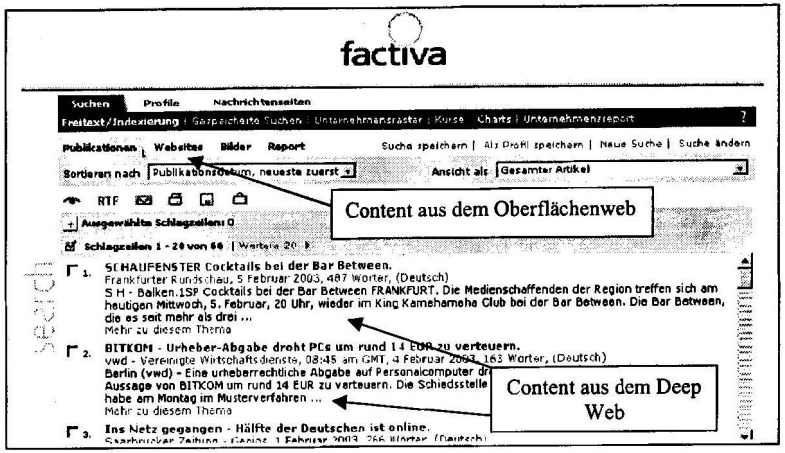

Abbildung 1-6: Factiva.com. Der Spezialist für proprietären Inhalt bietet als Querweltein-Ergänzung auch (unter „Websites") themenspezifische Quellen aus Websites an.

Dass an gewissen Stellen dann ein Kassenhäuschen auftritt, dürfte kaum schaden – der Nutzer entscheidet, ob ihm die volle Information Geld wert ist. Die Strategie der singulären Datenbankanbieter müsste sich eigentlich mit der der Suchmaschinen decken, da sie wohl daran interessiert sind, ihre Informationen soweit wie möglich zu streuen. Der große Vorteil fachspezifischer Datenbanken liegt darin, mit kontrolliertem Vokabular zu arbeiten und enge Nutzergruppen anzusprechen. Hier muss es Strategie sein, alle Informationsressourcen, also auch die des Oberflächenweb, mittels einheitlicher Methoden und Werkzeuge auszuwerten. Mit der Erkenntnis, die Welt der Informationen im Internet mit der Welt der Informationen, die via Internet erreichbar werden, querweltein zu verbinden, dürfte eine Umsetzung des gemeinsamen Zugriffs auf alle digitalen Online-Informationen realistisch erscheinen. Dass die Retrievalsysteme und insbesondere das Relevance Ranking vor neuen Herausforderungen stehen, ist klar, denn die Menge der Datensätze wird massiv erweitert.

Nur zwei Beispiele sollen die Informationsmengen verdeutlichen. Zu der Zeit, als dieses Buch geschrieben wird (im März 2003), hat die Datenbasis von Google 3,1 Milliarden Datensätze, d.h. Dokumente aus dem Oberflächenweb. Zur gleichen Zeit umfasst der Content-Aggregator Dialog Corp. über 1,2 Milliarden Datensätze, hier als

Teil des Deep Web zu wissenschaftlicher und wirtschaftsbezogener Literatur. Leider gibt es umfassende Suchsysteme zum Querweltein-Retrieval noch nicht. Unsere Beispiele zeigen denn auch nur erste Ansätze zur Annäherung der unterschiedlichen Welten. Der Suchende ist aufgefordert, selbst querweltein zu denken und zu suchen.

Ein Nutzer, der ein Informationsproblem lösen möchte, ist – je nach Fragetyp – mit einigen wenigen Datensätzen, im Extremfall mit genau einem Datensatz zufrieden, Hauptsache, das Problem wird in der Tat gelöst. Recherchen in den Welten des Internet kommen damit sehr nahe an das berühmte Suchen nach der Nadel im Heuhaufen. Um die „Informationsnadel" möglichst zielgenau zu finden, ist Wissen über das „Heu" (sprich: die Wissensbasen der Suchsysteme) sowie über die Strukturierung des „Haufens" (sprich: die Arbeitsweise der Retrievalalgorithmen) nötig. Und dies möchte das Buch leisten: Wir wollen dem Leser einen Einblick geben, wie die Systeme in den unterschiedlichen Teilwelten des Internet arbeiten. Und wir wollen dafür sensibilisieren, dass es überhaupt die Teilwelten gibt und dass der Nutzer einen Überblick über sie haben muss, will er die relevanten Informationen – und nur diese – finden.

Literatur zu Kapitel 1

Michael K. Bergman: The Deep Web: Surfacing Hidden Value. – In: The Journal of Electronic Publishing 7 (2001) Iss. 1 – URL: www.press.umich.edu/jep/07-01/bergman.html.

Chris Sherman; Gary Price: The Invisible Web. – Medford: Information Today, 2001.

Mechtild Stock; Wolfgang G. Stock: Internet-Suchwerkzeuge im Vergleich. Teil 1: Retrievaltest mit Known Item Searches. - In: Password Nr. 11 (2000), S. 23-30.

Das Buch unternimmt anhand von Fallbeispielen einen Streifzug durch die unterschiedlichen Weltregionen des Internet. Die Autoren zeigen jeweils, wie die Suchwerkzeuge und Datenbanken arbeiten, wie sie zu ihren Dokumenten kommen, wie diese inhaltlich ausgewertet werden, wie die Rangfolgen in den Ergebnislisten zustande kommen und wie die Benutzerschnittstellen aussehen. Dieses Hintergrundwissen hilft dem Nutzer, mit den Suchwerkzeugen und Datenbanken optimal zu arbeiten.

Das Buch ist so geschrieben, dass der Leser in recht kurzer Zeit einen Einblick in die Arbeitsweisen der Systeme im Web findet und – dank der detaillierten Fallbeispiele – dieses Wissen auch in der Praxis umsetzen kann.

Die Interessenten:
Zielgruppe sind Mitarbeiter von Unternehmen und anderen Institutionen, die beruflich im Internet recherchieren oder dies in Zukunft tun wollen und deren Arbeitszeit zu knapp ist, um umständlich mit Suchwerkzeugen zu experimentieren und danach trotzdem nur ein suboptimales Rechercheergebnis vorzulegen.

Die Autoren:
Mechtild Stock schreibt wissenschaftsjournalistische Artikel zu allen Bereichen der elektronischen Informationsdienste. Sie ist Mitarbeiterin bei „Password", dem deutschsprachigen Newsletter für information Professionals. Zudem ist sie Spezialistin für Systemtests von digitalen Firmeninformationen und von kostenpflichtigen Angeboten der kommerziellen Informationsanbieter.
Wolfgang G. Stock hat den Lehrstuhl für Informationswissenschaft der Heinrich-Heine-Universität Düsseldorf inne; seine hauptsächlichen Forschungs- und Lehrgebiete sind Wissensrepräsentation und Information Retrieval.

ISBN 3-8169-2278-3