Google Scholar

Aufbau und strategische Ausrichtung des Angebots sowie Auswirkungen auf andere Angebote im Bereich der wissenschaftlichen Suchmaschinen

Expertise im Auftrag des Hochschulbibliothekszentrums Nordrhein-Westfalen

Dirk Lewandowski
dirk.lewandowski@uni-duesseldorf.de

Düsseldorf, im Februar 2005

Inhaltsverzeichnis

0. Management Summary

Google, die weltweit erfolgreichste Suchmaschine, begibt sich mit seinem neuen Dienst Google Scholar auf das Feld der wissenschaftlichen Informationen. Der technische Aufbau dieses Angebots ist an die etablierte Websuche angelehnt und adaptiert deren Erfolgsfaktoren Bedienbarkeit, Geschwindigkeit, Indexgröße und Qualität des Rankings.

Neben der Websuche bietet Google weitere Dienste an, die unter technologischen Gesichtspunkten oder in Hinblick auf Erfahrungen innerhalb von Partnerschaften mit Dritten von Bedeutung für Google Scholar sind: Webverzeichnis, Google News, Newsgroups, Froogle (Produktsuche), Google Print (digitalisierte Bücher), Library Project (Digitalisierung von Bibliotheksbeständen) und Google Answers (kostenpflichtiger Auskunftsdienst).

Zu diesen Angeboten bestehen Synergien, die bisher jedoch noch nicht ausgeschöpft werden. Bei einem regulären Start des Angebots (bisher ist es nur in einer Beta-Version verfügbar) ist allerdings mit einer entsprechenden Ausnutzung zu rechnen.

Bisher werden in Google Scholar Inhalte aus dem freien Web und Inhalte von Verlagspartnern erfasst. Der Umfang liegt zwischen zwei und sieben Millionen Dokumenten. Es sind unterschiedliche Inhalte erfasst: Neben Artikeln aus Fachzeitschriften, die ein Peer-Review-Verfahren durchlaufen haben, sind auch Preprints, Reports u.ä. enthalten.

Die Dokumente sind im Volltext erschlossen; das teils in den Quellen verwendete kontrollierte Vokabular findet keine Berücksichtigung. Die Suche erfolgt über ein einziges Eingabefeld in der Standardsuche oder eine erweiterte Suchmaske. Die Suche nach Autoren oder Zeitschriftentiteln ist möglich, funktioniert aber nur unbefriedigend.

Die aus der Analyse des Angebots abgeleiteten Empfehlungen für ein Konkurrenzsystem lauten:

- Zusammenfassung der Repräsentationen der erschlossenen Dokumente in einer Datenbank anstatt einer Meta-Suche. Damit wird die Geschwindigkeit erhöht.
- Transparent machen, welche Quellen erschlossen werden und in welchem Umfang (bspw. welche Jahrgänge).
- Umfangreiches Quellenspektrum aus Bibliotheks-, Verlags-, Open-Access-Inhalten.
- Abdeckung des gesamten Fächerspektrums.
- Verwendung eines Rankings nach Relevanz, daneben wählbares Ranking nach anderen Faktoren wie z.B. dem Datum.
- Erschließung der Volltexte und Verwendung eines kontrollierten Vokabulars (bzw. mehrerer).
- Bei der Gestaltung der Suchoberfläche Orientierung an der Einfachheit des Google-Suchinterfaces, dazu Bedürfnisse der Profis berücksichtigen.
- Umfassende Suchmöglichkeiten sowohl über Formulare als auch über eine Abfragesprache.

- Neben der Suche auch Browsing ermöglichen.
- Anfallende Kosten schon in der Trefferliste ausweisen. Höchstmögliche Kostentransparenz.
- Verfügbarkeitstransparenz schaffen.
- Mehrwerte anbieten: Alert-Service, Erstellung von Bibliographielisten, o.ä.

Letztlich sollte die Zeit bis zur Einführung der „Vollversion“ von Google Scholar und dem damit zu erwartenden Popularitätsschub genutzt werden, um ein konkurrenzfähiges Angebot zu schaffen.

1. Hintergründe zur Stellung Googles

1.1. Suchmaschinen-Landschaft USA / Deutschland

Der Markt der Suchmaschinen beschränkt sich sowohl international als auch in Deutschland auf wenige Anbieter. Dies sind die Suchmaschinen Google, Yahoo und MSN (Microsoft). Zwar existiert eine größere Anzahl (auch bekannter und viel genutzter) Such*seiten*, diese beziehen ihre Ergebnisse jedoch meist entweder von Google oder Yahoo; Microsoft mit seiner noch neuen eigenständigen Lösung lizensiert bisher keine Treffer an andere Anbieter. Suchmaschinen, die eine eigene Datenbank mit Webseiten betreiben und technologisch eine gewisse Bedeutung haben (z.B. Ask Jeeves/Teoma), existieren zwar, haben jedoch international keine größere Bedeutung. Auf nationaler Ebene existieren teils eigenständige Suchmaschinen, aber auch deren Bedeutung ist als gering anzusehen.

Die großen internationalen Suchmaschinen unterscheiden sich in ihrer Ausrichtung deutlich. Die Suchmaschinen von Yahoo und Microsoft sind in die von den jeweiligen Unternehmen betriebenen Portalangebote integriert und bilden nur einen Teil eines Gesamtangebots (wenn auch einen von hoher Bedeutung) – anders ist dies bei Google: dieser Anbieter konzentriert sich klar auf die Suchfunktionalität. Allerdings ist auch hier in den letzten Jahren eine Diversifizierung festzustellen, die sich jedoch (zumindest bisher) nicht in der Gestaltung des Webangebots niedergeschlagen hat. Die Elemente des Google-Angebots werden in Abschnitt 1.3 ausführlicher dargestellt.

Die Reduzierung des Suchmaschinen-Angebots auf nur wenige Anbieter hat zur Folge, dass auch bei den Suchergebnissen keine Vielfalt besteht. Insbesondere der Nachweis auf den vorderen Listenplätzen bei Google ist sowohl für Unternehmen als auch für Institutionen von zentraler Bedeutung und entscheidet oft über wirtschaftlichen Erfolg oder Misserfolg. Daher sind die ersten Trefferplätze für viele Suchbegriffe heiß umkämpft, wodurch wiederum eine eigene Branche (Search Engine Optimizers, SEO) entstanden ist, die Webangebote hinsichtlich ihrer Sichtbarkeit in Suchmaschinen optimiert. Die angebotenen Leistungen reichen (je nach Anbieter) von einer Beratung der Kunden hinsichtlich Inhalt und Gestaltung der Websites bis zu (von den Suchmaschinen in ihren Nutzungsbedingungen verbotenen) Manipulationen. Seit einigen Jahren ist so in den Trefferlisten der Suchmaschinen ein Primat des Kommerziellen zu verzeichnen: Bei Suchbegriffen, die auf eine kommerzielles Interesse hindeuten *können* (z.B. „Urlaub“), werden bevorzugt „optimierte“ Webseiten von kommerziellen Anbietern angezeigt.

Auf der Nutzerseite wird diese Problematik nicht erkannt, wie insgesamt die Nutzer nur wenig Kenntnisse über den Aufbau, die Recherchefunktionen, das Ranking der Trefferlisten und die Finanzierung von Suchmaschinen haben. Auch hierbei handelt es sich um ein internationales Phänomen; auch in den hinsichtlich der Suchmaschinennutzung als erfahrener anzusehenden USA können die Nutzer nicht nachweisbar besser mit Suchmaschinen umgehen, noch wissen sie mehr über deren Hintergründe.

Aus dem Nutzerverhalten lassen sich allgemeine Orientierungen für Suchmaschinen ableiten: Sie müssen sich auf diese Nutzer fokussieren und deren Rechercheverhalten akzeptieren, letztlich durch übersichtliche Suchinterfaces ohne erweiterte Funktionen sogar unterstützen. Während die kommerziellen Suchmaschinen dieses verinnerlicht haben und alle erweiterten Suchfunktionen in eine eigens dafür vorgesehene „erweiterte Suche“ (die jedoch von der Nutzerschaft nur schlecht angenommen wird) verbannt haben, sind öffentliche Angebote oft nicht an dieses Nutzerverhalten angepasst. Hier ist eine Umorientierung erforderlich. Es ist davon auszugehen, dass sich auch im wissenschaftlichen Kontext das Rechercheverhalten bei einem Großteil der Nutzer nur wenig von dem der Laien, die in den kommerziellen Suchmaschinen recherchieren, unterscheidet.

Letztlich ist die „Gatekeeper-Funktion“ der Suchmaschinen zu betonen: Sie entscheiden, welche Informationen genutzt werden und welche in der Menge der Treffer „untergehen“. Dass die Entscheidung über den Wert der Treffer algorithmisch (und damit vordergründig „objektiv“) erfolgt, ändert nichts daran, dass eine Entscheidung über den Wert der (möglichen) Treffer erfolgt. Neben der Problematik der geringen Zahl der existierenden bzw. genutzten Suchmaschinen ergibt sich ein weiteres Problem daraus, dass sich die Suchergebnisse der großen Suchmaschinen zumindest bei populären Suchanfragen stark ähneln. Die allen großen Suchmaschinen zugrunde liegende Ableitung eines Qualitätsurteils aus der Auswertung der Verlinkungsstruktur (s.u.) sorgt dafür, dass bestimmte Arten von Dokumenten in den Trefferlisten bevorzugt werden. Allerdings regt sich in Fachkreisen zunehmend Kritik an der Qualitätsbestimmung allein auf der Basis solcher linktopologischer Verfahren.

Auch gegen die Monopol- bzw. Oligopolbildung auf dem Suchmaschinenmarkt haben sich Initiativen gebildet, an erster Stelle ist hier in Deutschland der SuMa e.V. zu nennen, der sich für eine verteilte Suchmaschinenstruktur, die auf Basis von Metasuchmaschinen zusammengefasst werden soll, stark macht. Allerdings ist auch diese Initiative kritisch zu sehen; erfolgversprechend scheint eher eine Stärkung einerseits der Suchmaschinen-Forschung, andererseits der kleineren Anbieter. Sowohl Suchmaschinenforschung als auch –entwicklung finden bisher weitgehend in den USA statt. Weiterhin ist eine Tendenz zu verzeichnen, dass die Ergebnisse der zu einem großen Teil bei den kommerziellen Anbietern stattfindenden F&E nicht mehr in großem Umfang publiziert werden.

Während also der Markt für die reguläre Websuche fest in kommerzieller Hand ist, gab es bislang keine wissenschaftlichen Suchmaschinen, die eine dominierende Stellung einnahmen. Zwar wurden Suchsysteme durchaus auch von kommerziellen Anbietern entwickelt, die Kontrolle über den Zugang zu diesen Systemen unterlag aber in der Regel den öffentlichen Einrichtungen, die diese Systeme lizensierten. Daneben existierten öffentlich erstellte und finanzierte Systeme.

Mit dem Eintritt von Google in den Markt für wissenschaftliche Informationen werden sich Verschiebungen auf diesem Markt ergeben. Die Hintergründe zu Google Scholar und seiner (angestrebten) Stellung in diesem Markt werden im Folgenden dargestellt.

1.2. Technische Hintergründe Google

Der relativ schnelle und durchschlagende Erfolg von Google gründet sich im Wesentlichen auf vier Faktoren: Bedienbarkeit, Geschwindigkeit, Indexgröße und die Qualität des Rankings.

Um die große Bedeutung dieser Faktoren zu verstehen, ist ein Blick auf die Suchmaschinen-Landschaft des Jahres 1998 nötig. Damals orientierten sich Suchmaschinen weitgehend an klassischen Information-Retrieval-Systemen, d.h. das Ranking fand auf rein textstatistischer Ebene statt. Die Grundannahme lautet hier, dass die Bedeutung eines Dokuments für einen Suchbegriff von der Häufigkeit des Vorkommens dieses Begriffs im Dokument sowie von der Stellung des Begriffs im Dokument abhängt. Dies bedeutet, dass im Ranking solche Dokumente, in denen der Suchbegriff häufig und an exponierten Stellen (bspw. in der Überschrift und am Anfang des Dokuments) vorkommt, bevorzugt werden. Es ist leicht verständlich, dass solche Verfahren mit nur geringem Aufwand manipuliert werden können: Man muss nur die gewünschten Begriffe im Dokument häufen und an entsprechender Stelle eintragen. Dies führte so weit, dass die Suchmaschinen sich größtenteils nicht einmal selbst bei der Eingabe ihres Markennamens an erster Stelle anzeigten, sondern eine andere Seite, die eine höhere Keyword-Dichte aufwies.

Natürlich versuchten die Suchmaschinen-Betreiber, solche Manipulationen zu unterbinden, indem beispielsweise (auch heute noch eingesetzte) Quoten errechnet wurden, die besagen, wie hoch der Anteil von Keywords in Dokumenten bei einer „natürlichen Seitenerstellung“ sein darf. Wird dieser Wert überschritten, steht das Dokument im Verdacht, manipuliert zu sein.

Allein auf klassischen Verfahren basierende Suchmaschinen hatten weiterhin das Problem, dass sie nicht zwischen dem Original und einer Kopie eines Dokuments unterscheiden konnten. Um auf die vorderen Plätze der Trefferlisten zu gelangen, reichte es aus, eine bereits gut gerankte Seite zu kopieren und der Suchmaschine zur Indexierung anzubieten. Die auf diese Seite gelenkten Besucher konnte man dann durch die Veränderung der Links auf eigene Angebote „umleiten“.

Die von Google in diesem Umfeld eingeführte Neuerung im Ranking war ein Rankingfaktor, der sich nicht auf deren Inhalt, sondern auf die Verlinkungsstruktur der Dokumente bezog: das sog. PageRank. PageRank ist ein statischer Faktor, der jedem von Google indexierten Dokument zugewiesen wird und eine Qualitätsbewertung darstellt. Die Grundannahme ist aus der Bibliometrie bzw. Zitationsanalyse übernommen: Ein Dokument, welches häufig zitiert wird, ist wohl bedeutender als eines, das weniger häufig oder gar nicht zitiert wird. Für die Zwecke von Suchmaschinen wird jeder Link auf ein Dokument als Zitation angesehen bzw. als „Stimme für dieses“. Allerdings bleibt es nicht bei der reinen Link-Zählung, sondern die Links werden wiederum nach dem PageRank der Seiten bewertet, von denen sie ausgehen. Damit wird gewährleistet, dass ein Link von einer bedeutenden Seite (bspw. von der Yahoo-Homepage) als bedeutender gewichtet wird als der Link von einer unbedeutenden Seite (bspw. einer privaten Homepage). Die PageRank-Werte werden in einem iterativen Verfahren berechnet und unabhängig von einer späteren

Suchanfrage festgelegt. Das Ranking erfolgt später in durch ein Mischung aus textstatistischen Verfahren und dem PageRank.

Mit der Einführung des PageRank konnte Google die oben beschriebenen Probleme bewältigen. Es war nun auch möglich, zwischen dem Original und der Kopie einer Seite zu unterscheiden, weil eine Originalseite in aller Regel besser verlinkt ist als die Kopie. Durch die zusätzliche Auswertung der Ankertexte auf den verlinkenden Seiten wurde auch die Dokumentbeschreibung erweitert, so dass auch Dokumente, die selbst den gesuchten Begriff nicht in hoher Dichte oder sogar gar nicht enthielten, für eine Suchanfrage als relevant gewertet werden konnten.

Auch in Bezug auf die Menge der erschlossenen Dokumente war Google wenn nicht führend, so doch bald unter den großen Anbietern. Dabei machte sich die Suchmaschine auch einen Trick zunutze, indem auch Dokumente in den Index aufgenommen wurden, die (noch) gar nicht erfasst waren, die aber durch von anderen Dokumenten ausgehenden Links auf diese bekannt waren. Zur Erschließung wurden wiederum die Ankertexte verwendet. Zeitweise bestand etwa ein Drittel des Google-Datenbestands aus solchen „indirekt“ indexierten Seiten; auch heute noch finden sich solche Dokumente im Datenbestand. Inzwischen bietet Google den größten Datenbestand unter den Web-Suchmaschinen.

Ein weiterer wichtiger Faktor für den Erfolg von Google war die einfache Bedienbarkeit und die Übersichtlichkeit der Suchseiten. Während die anderen großen Anbieter zum Zeitpunkt des Markteintritts von Google gerade ihre Suchmaschinen zu Portalen mit umfangreichen Funktionen ausgebaut hatten und versuchten, die Nutzerströme in Richtung von in das Portal intergrierten kommerziellen Angeboten zu lenken, setzte Google schon zu Beginn auf das auch heute noch bekannte schlichte Interface, welches aus nur einem Suchfeld ohne viel „Drumherum“ besteht. Auch die Trefferlisten wurden in einem schlichten, übersichtlichen Layout präsentiert. Während inzwischen alle Suchmaschinen diesem Beispiel gefolgt sind und die Nutzer eine solche Ergebnispräsentation voraussetzen, war dies damals ein Alleinstellungsmerkmal.

Als letzter wichtiger Erfolgsfaktor ist die Geschwindigkeit von Google zu nennen. Auch hier fallen heute keine signifikanten Unterschiede mehr zu den anderen Suchmaschinen auf; zum Zeitpunkt des Starts von Google wurden die Trefferlisten der anderen Suchmaschinen aber einerseits aufgrund ihrer Überfrachtung mit weiteren Angeboten (und vor allem grafischer Werbung), andererseits aufgrund der Rechnerstruktur langsamer geladen. Zu den Neuentwicklungen von Google gehört der Aufbau einer Suchmaschine auf Basis eines Clusters aus handelsüblichen PCs mit dem freien Betriebssystem Linux. Gegenüber der damals üblichen Großrechnerarchitektur der Suchmaschinen bietet dieses System Vorteile im Preis und in der Skalierbarkeit. Das System kann leicht den benötigten Ressourcen entsprechend angepasst werden.

Heute sind die genannten Erfolgsfaktoren differenziert zu beurteilen. Andere kommerzielle Anbieter bieten inzwischen eine vergleichbare Qualität der Trefferlisten bei ähnlicher Geschwindigkeit. Die Frage der Geschwindigkeit ist inzwischen auch

zunehmend in den Hintergrund gerückt, da die Nutzer zunehmend über schnellere Verbindungen verfügen und allein dadurch auch komplexere Seiten in vertretbarer Geschwindigkeit geladen werden.

In Bezug auf die Indexgrößen gibt es keine aktuelle Untersuchung, die zuverlässige Zahlen liefert. Daher ist man weitgehend auf die Angaben der Betreiber selbst angewiesen; in vergangenen Untersuchungen konnte jedoch gezeigt werden, dass diese Angaben weitgehend zutreffend sind. Google gibt die Größe seines Index mit acht Milliarden Dokumenten an, Microsoft hat ca. fünf Milliarden Dokumente indexiert. Von Yahoo werden keine Angaben gemacht, aber auch dieser Index dürfte bei etwa fünf Milliarden Dokumenten liegen.

Hinsichtlich der Usability der Suchinterfaces bietet Google weiterhin als einzige der großen Suchmaschinen eine schlicht gestaltete Startseite, die die Suche in der Vordergrund stellt. Sowohl MSN als auch Yahoo haben die Suche in ihre Portalangebote eingebunden, allerdings bieten beide auch ein eigenes „Nur-Suche"-Interface an.

Der Erfolg von Google heute basiert in erster Linie auf den für die Anfangsjahre bedeutenden, oben beschriebenen Faktoren. Die Überlegenheit aus dieser Zeit wurde von dem Unternehmen geschickt genutzt und vom Marketing wurde das Image der Überlegenheit Googles weitergetragen und so letztlich eine der weltweit stärksten Marken aufgebaut. Der Erfolg beruht heute als weitgehend auf einer von den Nutzern angenommenen Überlegenheit dieser Suchmaschinen unabhängig von einer empirischen Evidenz.

Ein Faktor, der Google einerseits im Gespräch hält, andererseits das Image als technologisch führendes Suchmaschinen-Unternehmen unterstützt, ist das Anbieten immer neuer Leistungen und Erweiterungen des bestehenden Angebots. Neben der regulären Websuche werden mittlerweile unter anderem eine Nachrichtensuche, eine lokalisierte Suche und eine Desktop-Suche angeboten. Google präsentiert sich damit als innovatives Unternehmen, das neue Anwendungen in seinen „Google Labs" (labs.google.com) öffentlich testet. Die von Google angeboten Neuerungen werden in der Öffentlichkeit besonders positiv wahrgenommen; den anderen Anbietern gelingt es nicht, sich als ähnlich innovativ zu platzieren, unabhängig davon, welche Innovationskraft ihre Angebote tatsächlich haben bzw. ob die Angebote von Google gegenüber denen der Konkurrenz tatsächlich bahnbrechende Neuerungen bringen.

1.3. Contentstrategie der Firma Google

Google hat frühzeitig erkannt, dass sich die Informationserschließung durch Suchmaschinen nicht allein auf das „reguläre Web" (also das „surface Web", der Teil des WWW also, der durch linkbasiertes Crawling von den Suchmaschinen erfasst werden kann) beschränken sollte und deshalb das Angebot stetig um weitere Inhalte erweitert. Diese Ausrichtung spiegelt sich in einem der Firmenmottos wieder: „To organize the world's information and make it universally accessible and useful." Dieses

Ziel mag angesichts der unglaublichen Menge der weltweit verfügbaren Informationen als zu hoch gesteckt oder gar vermessen erscheinen, es zeigt jedoch den universellen Anspruch, der verfolgt wird.

Im Folgenden werden die wichtigsten über Google recherchierbaren Datenbestände kurz beschrieben, wobei diejenigen Angebote herausgegriffen werden, die in einem wissenschaftlichen Kontext eine Rolle spielen (können) bzw. Technologien einsetzen, die für den Aufbau von Google Scholar relevant sind.

Verzeichnis: Google unterhält kein eigenes Verzeichnis, sondern verwendet die frei verfügbaren Daten des Open Directory Project (ODP). Im Unterschied zu anderen ODP-basierten Verzeichnissen werden die Treffer innerhalb der Kategorien jedoch nicht alphabetisch, sondern nach ihrer Relevanz (beruhend auf dem PageRank-Wert), angeordnet.

Google News: Dieses Angebot aggregiert Nachrichten aus einer Vielzahl von Nachrichtenquellen aus dem Web. Neben der Suchmöglichkeit wird eine automatisch erstellte Nachrichtenschau angeboten, die die als besonders wichtig angesehenen aktuellen Meldungen aus den unterschiedlichen Rubriken auf einer Seite zusammenführt. Zu betonen ist, dass Google News kein vollständiges Pressearchiv liefert, sondern nur die Dokumente indexieren kann, die von den Anbietern kostenlos im Netz angeboten werden. Die Nachrichten bleiben auch nur bis zu vier Wochen im Bestand, ältere Nachrichten können nur noch über die reguläre Web-Suche recherchiert werden. Eine Besonderheit ist die Indexierung von Nachrichten, die über das Angebot der Nachrichtenquelle selbst nicht ohne vorherige Registrierung erreicht werden können. Teils bestehen Vereinbarungen zwischen Google und den Betreibern der Nachrichtenangebote, die den Google-Crawlern Zugang zu den Nachrichteninhalten gewähren. Diese Form der Kooperation gelangt bei den in Kapitel 3 behandelten wissenschaftlichen Inhalten zu einer noch größeren Bedeutung.

Newsgroups: Das Archiv stellt die weltweit größte Sammlung von Newsgroup-Beiträgen dar.

Froogle ist eine Produktsuchmaschine, die die Recherche nach Anbietern eines Produkts und einen Preisvergleich unterschiedlicher Anbieter möglich macht. Für den Kontext dieser Untersuchung ist die hier stattfindende Zusammenarbeit mit den Produktanbietern von Bedeutung. Der Datenbestand von Froogle besteht nämlich einerseits (zum geringeren Teil) aus Informationen, die aus gecrawlten Seiten des öffentlichen Web extrahiert wurden, andererseits aus von den Anbietern zu Google überspielten *feeds*, also Datenströmen mit den Produktinformationen.

Google Print. In diesem Projekt werden die Inhalte von Büchern und Zeitschriften erschlossen. Die Bücher wurden eingescannt und der Nutzer bekommt auf eine Anfrage in der regulären Web-Suche, für die auch Bücher-Ergebnisse gefunden werden, einen Hinweis auf diese Treffer oberhalb der Trefferliste angezeigt (allerdings nur in der US-Version von Google). Eine eigene Suchseite für Google Print existiert nicht. Die Print-Treffer umfassen neben bibliographischen Angaben und Links auf Online-Buchhändler auch die Möglichkeit, sich Seiten des entsprechenden Buchs im

Original-Layout anzeigen zu lassen. Angezeigt wird die Seite, die als für die eingegebenen Suchbegriffe als besonders relevant angesehen wird. Der Nutzer hat von dort aus die Möglichkeit, jeweils maximal zwei Seiten vor und zurück zu blättern. Zusätzlich ist es möglich, das Inhaltsverzeichnis und wo vorhanden das Register anzusehen. Die erfassten Zeitschriftenartikel liegen im Volltext vor. Weder zur Zahl der erfassten Bücher noch der der Zeitschriftenaufsätze oder deren Quellen werden von Google Angaben gemacht.

Library Project. Im Rahmen dieses Projekts werden Bibliotheksbestände großer amerikanischer Universitätsbibliotheken in großem Umfang eingescannt. In den kommenden Jahren sollen mehrere Millionen Bände erfasst werden; damit dürfte dieses Projekt das weltweit größte Digitalisierungsprojekt sein. Die Vereinbarungen mit den Bibliotheken sind nicht auf eine bestimmte Anzahl von Büchern oder auf bestimmte Bestände festgelegt, sondern potentiell auf den gesamten Bestand angelegt.

Besonders betont werden muss in diesem Kontext nochmals, dass hier ein privatwirtschaftlicher Anbieter einen Dienst aufbaut, der traditionell eher dem öffentlichen (Bibliotheks-)Bereich zuzuordnen ist.

Google Answers ist ein kommerzieller Recherchedienst, der auf dem Bieterprinzip beruht. Auskunftsfragen können gestellt werden, wobei ein Betrag angegeben wird, den der Fragende bereit ist, für die Auskunft zu bezahlen. Zur Beantwortung berechtigt sind Personen, die sich vorher bei Google als Rechercheure angemeldet haben und einen (allerdings recht anspruchslosen) Recherchetest bestanden haben. Der aus der Beantwortung der Frage eingenommene Betrag wird zwischen Google und dem Rechercheur aufgeteilt. Google Answers wurde von der Fachöffentlichkeit weitgehend abgelehnt, da die Rechercheure mit der Qualität der Arbeit von Information Professionals nicht mithalten können und die gebotenen Preis zu gering sind, um eine professionelle Beantwortung gewinnbringend durchzuführen. Allerdings wurde diesem Angebot von bibliothekarischer Seite nichts Vergleichbares entgegengesetzt. Google Answers ist allerdings auch nicht als großer Erfolg zu werten. Zwar werden stetig Fragen gestellt und beantwortet, das Volumen ist jedoch mit etwa 100 bis 120 Anfragen pro Tag als relativ gering zu bewerten, zumal für ein potentiell weltweites Angebot. Die bisher noch starke US-Zentrierung soll aufgehoben werden, indem das Angebot noch in diesem Jahr auf Deutschland ausgeweitet werden soll. Damit ergibt sich eine direkte Konkurrenz vor allem für die Online-Auskunftsdienste der Bibliotheken.

Die dargestellten Angebote zeigen einerseits die zunehmende Diversifizierung der Inhalte innerhalb des Google-Angebots, andererseits die Potentiale für eine wissenschaftliche Suchmaschine. Google konnte im Lauf der Jahre Erfahrungen mit der Einbindung fremder Inhalte sammeln, dazu kommen Erfahrungen mit der Partnerschaft mit Fremdanbietern (allerdings weniger auf einer kommerziellen Ebene als auf der Inhaltsebene) und die Erfahrungen aus einer „kollaborativen Informationsvermittlung“.

Als problematisch kann die bisherige Form der Integration der unterschiedlichen Datenbestände angesehen werden. Zwar sind einige Bestände direkt über *tabs* oberhalb des Eingabefelds zu erreichen; andere Angebote sind allerdings wenig prominent platziert und werden nur erreicht, wenn sie gezielt angesteuert werden; d.h. in der Regel wird hier die Kenntnis des entsprechenden Angebots bereits vorausgesetzt. Für die Zukunft ist jedoch mit einer verbesserten Integration der Angebote zu rechnen, so dass Google (wie andere Suchmaschinen auch) noch stärker nutzerzentriert aufgebaut sein wird. Auf die Eingabe einer Suchanfrage in das Suchfeld auf der Google-Startseite können zwar schon wie bisher auch direkt Ergebnisse angezeigt werden, allerdings könnten diese (mehr als bisher) um Hinweise auf weitere Dokumentkollektionen ergänzt werden, z.B. auch mit einer Ergänzung dieser Hinweise um die Anzahl der Treffer in dem jeweiligen Bestand.

2. Google Scholar: Technik und Einbindung in das Google-Angebot

2.1. Aufbau und technische Hintergründe Google Scholar

Google greift für das Scholar-Angebot auf die technische Plattform seiner Web-Suchmaschine zurück, wobei die Verfahren der Erschließung und des Rankings an die Besonderheiten der wissenschaftlichen Inhalte angepasst und entsprechend erweitert werden. Der Rückgriff auf die „traditionelle“ Google-Technik im Hintergrund bringt die Vorteile der für den Nutzer gewohnten Rechercheoberfläche sowie die oben beschriebenen Geschwindigkeitsvorteile mit sich.

Das Interface ist dem gewohnten Google-„look and feel“ angepasst und bietet wie die Websuche auch ein einziges Eingabefeld, in das die Suchanfragen eingetragen werden. Solche Interfaces sind grundsätzlich eher für „quick and dirty“-Recherchen ausgelegt, obwohl sie bei Kenntnis der entsprechenden Abfragesprachen auch für komplexe Anfragen genutzt werden können. Für diese Anfragen steht außerdem ein erweitertes Suchformular zur Verfügung, welches die Möglichkeit zur Einschränkung von Suchanfragen bietet.

Das Ranking in Google Scholar basiert auf den in der Websuche eingesetzten Rankingverfahren, wobei die Gewichtungen entsprechend der wissenschaftlichen Zwecke angepasst wurden; allerdings ist auch hier das Rankingverfahren nicht veröffentlicht und könnte auch nur mit großem Aufwand in Form eines *reverse engineering* annäherungsweise ermittelt werden.

Neben den üblichen informationsstatistischen Verfahren setzt Google Scholar auf eine Zitationsanalyse, die im nächsten Abschnitt beschrieben wird. Inwieweit auch weitere, durch die Indexierung in Google bereits bekannte Popularitätswerte der erschlossenen Seiten bzw. Sites mit in das Ranking eingehen, kann nicht bestimmt werden.

2.2. Zitationsanalyse

Für die in Google Scholar eingesetzte Zitationsanalyse kann Google einerseits auf das gut erforschte Feld der wissenschaftlichen Zitationsanalysen (Stichworte Science Citation Index, Eugene Garfield) zurückgreifen, andererseits auf die eigenen Erfahrungen mit dem Einsatz der Linkanalyse mit PageRank. Dazu kommen die Erfahrungen mit der wissenschaftlichen Suchmaschine CiteSeer (bzw. Researchindex), welche im Rahmen der Indexierung von Informatik-Papers eine automatische Extraktion von Zitationen einsetzt. Der neben Lee Giles zweite „Vater“ von Citeseer, Steve Lawrence, ist mittlerweile bei Google beschäftigt. Man kann auch davon sprechen, dass Google Scholar nur das CiteSeer-System auf seiner (perfomanteren) Plattform umsetzt und auf alle Wissenschaftsgebiete erweitert.

Unter jedem in Google Scholar angezeigten Treffer wird die Zahl der Zitationen angegeben. Auch hier gilt das bereits in der regulären Google-Suche verwendete Prinzip, dass eine Zitation als eine Stimme für das zitierte Dokument angesehen wird.

Bei den Recherchetests fiel auf, dass Dokumente allerdings recht häufig mehrfach in den Trefferlisten auftauchen und damit auch deren Zitation gesondert gezählt werden.

Analog zur Websuche werden in den Trefferlisten auch Dokumente gezeigt, die nicht im Datenbestand von Google Scholar vorhanden sind; dies sind in erster Linie Bücher. Diese Eintragungen enthalten nur die Angaben zu Autoren und Titel, dazu kommen Links auf Online-Buchhändler und teils auf den World Cat des OCLC. In einem nächsten Schritt kann dann herausgefunden werden, ob das Werk in einer Bibliothek in der Nähe des Nutzers vorhanden ist. Dies funktioniert bisher nur für US-Bibliotheken; mit einer entsprechenden Einbindung auch deutscher Bibliotheken ist bei einem offiziellen Start des Angebots zu rechnen.

2.3. Einbindung von Google Scholar in das Gesamtangebot google.com

Wie auch bei anderen neuen Angeboten auf der Google-Plattform üblich, befindet sich auch Google Scholar im Betastadium. Google ist dafür bekannt, dass neue Angebote lange in diesem Status verharren, bis sie dann als fertige Version freigegeben werden. So befindet sich das im Oktober 2002estartete Angebot Google News beispielsweise immer noch im Beta-Stadium, obwohl es inzwischen durchaus als etabliert betrachtet werden kann. Den Beta-Angeboten eigen ist, dass sie zum großen Teil innerhalb der Google-Website nicht prominent platziert sind, sondern nur entweder über die Übersichtsseite der *Google Labs* (des „Experimentierfelds“ von Google) oder über eine eigene URL erreicht werden können. Die Angebote werden auch in keiner Form beworben; allerdings entstand gerade beim Start von Google Scholar ein großes Interesse seitens der Presse, so dass davon ausgegangen werden kann, dass das Angebot innerhalb der Scientific Community schon eine relativ hohe Bekanntheit erreicht hat. Nutzungszahlen liegen wie bei solchen Einzelangeboten innerhalb von Websites üblich nicht vor.

Bei einem offiziellen Start von Google Scholar und einer besseren Einbindung in das Gesamtangebot von Google (etwa durch eine Ankündigung auf der Startseite der regulären Web-Suche bzw. einen eigenen Link in einem *tab* oberhalb des Suchfelds) dürfte die Popularität des Angebots noch wesentlich steigen. Google verfügt mit seinem sehr stark frequentierten Angebot über die Möglichkeit, Neuentwicklungen rasch einem großen Publikum bekannt zu machen. Ein weiterer Faktor für die Steigerung der Nutzungszahlen stellen die Anpassungen an die unterschiedlichen Länderseiten bzw. die Erstellung unterschiedlicher Sprachversionen dar. Bisher ist Google Scholar (wie auch die anderen *Labs*-Angebote) nur in einer einzigen Version verfügbar; auch die Inhalte haben ihren Schwerpunkt noch deutlich bei englischsprachigen Dokumenten (s.a. Abschnitt 3.1). Zwar liegen die Steigerungspotentiale bei Google Scholar aufgrund der Internationalität der Wissenschaft geringer als bei anderen Beta-Angeboten, eine deutliche Steigerung dürfte sich aber vor allem durch eine Ausweitung der Inhalte ergeben.

2.4. Mögliche Synergien mit weiteren Google-Angeboten

Besonders bei den Angeboten Google Print und dem Library Project ist mit einer weiteren Einbindung in Google Scholar zu rechnen. Aber auch zur Websuche, dem Verzeichnis und den in die Websuche eingebundenen Definitionen bestehen Synergien, die sich für die Verbesserung des Angebots ausnutzen lassen.

Die in Google Print enthaltenen bibliographischen Informationen lassen sich gut als Ergänzung bzw. zur Überprüfung der in Google Scholar enthaltenen, aus den Zitationen gewonnenen Angaben verwenden. Die Qualität dieser von den Verlagen bereitgestellten Informationen liegt weit höher als die der Zitationen. Mit einer Erweiterung der Angaben innerhalb des Print-Angebots ist auf längere Frist zu rechnen, zumal Amazon ein entsprechendes Konkurrenzprodukt anbietet. Weiterhin interessant ist, dass Amazon auch eine eigene Suchmaschine (A9.com) betreibt, die eine parallele Suche im Google-Web-Index und der Amazon-Bücherdatenbank erlaubt.

Ähnliches gilt auch für das Library Project. Die eingescannten wissenschaftlichen Bücher könnten in Google Scholar eingebunden werden. Sie wären im Volltext recherchierbar und könnten nach Bedarf am Bildschirm gelesen oder ausgedruckt werden, sofern dem keine rechtlichen Fragen entgegenstehen. Ähnlich wie bei Google Print und bei Amazon könnte nur ein Ausschnitt je Suchanfrage verfügbar gemacht werden. Ebenso ist (wie heute schon bei den kostenpflichtigen Aufsätzen in Google Scholar) eine Bezahllösung denkbar, die dann die vollständigen Bücher verfügbar macht.

Für traditionelle Bibliotheksangebote besteht hier die Gefahr, dass Teile des vorgehaltenen Bestands aufgrund des Google-Angebots nicht mehr genutzt werden. Weiterhin wird die Diskussion um den lokal vorhandenen Bestand wieder aufflammen; welche der digital vorhandenen Bücher sollen in der lokalen Bibliothek überhaupt noch in Printform vorgehalten werden?

Auch wenn die Bibliotheken in der Lage wären, elektronische Bücher, die über Google Scholar nur kostenpflichtig oder in Ausschnitten genutzt werden können, ihren Nutzern kostenfrei anzubieten, stellt sich die Frage, ob Benutzer bei einer fairen Preisgestaltung nicht bereit wären, für die Nutzung der Werke zu bezahlen. Sie dies, weil sie keine Kenntnis der entsprechenden Bibliotheksangebote haben oder den Weg über das Bibliotheksangebot als zu umständlich empfinden.

Die Einbindung der Definitionen in Google Scholar könnte analog zur bisherigen Einbindung in der Websuche erfolgen. Für Anfragen aus unterschiedlichen Fachgebieten könnten entsprechende Fachlexika hinterlegt werden, um den Ansprüchen der Wissenschaftler zu genügen.

Die Einbindung von Web-Verzeichnissen könnte Hinweise auf relevante Kategorien des Verzeichnisses bringen. Allerdings ist eine Einbindung des bisher verwendeten Verzeichnisses (ODP) in Hinblick auf dessen Qualität und Universalität als kritisch zu sehen. Hier ergäbe sich die Chance für bibliothekarisch betreute Linksammlungen, diese in Google Scholar einbinden zu lassen.

Die Einbindung in die Websuche schließlich könnte auf der Ebene der *tabs* und auf der Ebene des Hinweises bei einer Websuche erfolgen. Möglich wäre also einerseits die Möglichkeit, Google Scholar schon auf der Google-Startseits mittels eines Reiters auszuwählen (wie heute schon News, Bilder, usw.), andererseits könnte ein Hinweis auf Scholar-Treffer oberhalb der Trefferliste angezeigt werden, wenn in diesem Angebot Treffer gefunden werden. Solche Hinweise gibt Google schon auf das News- und das Bilder-Angebot.

3. Content

3.1. Art und Umfang der Inhalte

Das Angebot von Google Scholar ist nicht nach Publikationsformen getrennt und enthält neben Aufsätzen aus Zeitschriften und Tagungsbänden, die ein Peer-Review-Verfahren durchlaufen haben, auch Qualifikationsarbeiten, Bücher, Preprints, technische Reports und ähnliches.

Thematisch unterliegt das Google-Angebot keinen Beschränkungen; prinzipiell sollen alle Wissenschaftsdisziplinen erfasst werden. Erste Beobachtungen zeigen jedoch einen Schwerpunkt bei den Naturwissenschaften und der Technik gegenüber den Gesellschafts- und Geisteswissenschaften. Empirische Befunde hierzu, die auf einer validen Datenbasis beruhen, liegen allerdings noch nicht vor.

Die in Google Scholar erschlossenen Dokumente stammen sowohl aus dem offenen Web als auch aus proprietären Quellen von Wissenschaftsverlagen und Fachgesellschaften. Weiterhin ist eine Einbindung von Open-Access-Archiven vorgesehen, bisher konnte aber noch keine solche Einbindung festgestellt werden. Das Quellenspektrum wird von Google Scholar nicht bekannt gegeben, so dass eine Überprüfung der Quellen, die tatsächlich erschlossen werden, schwierig ist. Dies gilt besonders auch für die aus dem freien Web entnommenen Dokumente: Wie hier entschieden wurde, welche Dokumente wissenschaftlich sind und welche nicht, bleibt unklar. Eine reine Beschränkung auf wissenschaftliche Server wie Hochschulen und Forschungseinrichtungen kommt nicht in Frage, da diese in einem hohen Maß auch Dokumente enthalten, die nicht der von Google Scholar angestrebten Forschungsliteratur zugehörig sind (Pressemitteilungen, Selbstdarstellungen, biographische Informationen, usw.). Die Zuordnung der Dokumente zum Themenfeld Wissenschaft ist (zumindest bisher) zum Teil unzuverlässig. So fanden sich bei ersten Tests beispielsweise Romane im Bestand.

Von größerer Bedeutung dürfte allerdings die Durchbrechung der klassischen Trennung von wissenschaftlichen Werken nach ihrer Qualität/Autorität sein: Google Scholar durchmischt in den Trefferlisten alle als wissenschaftlich angesehenen Publikationsformen und gibt keine Hinweise darauf, ob eine vorherige Qualitätskontrolle wie etwa ein Peer Review stattgefunden hat. Dies führt so weit, dass in den Trefferlisten – vor allem bei der Suche mit deutschsprachigen Begriffen – häufig studentische Arbeiten auftauchen. Diesen soll hier nicht die Qualität grundsätzlich abgesprochen werden, mit einem Aufsatz in einer renommierten Fachzeitschrift sind jedoch wohl kaum zu vergleichen. Google Scholar birgt damit die Gefahr, dass Studenten (und auch Wissenschaftler) bei ihren Literaturrecherchen die Qualität der Dokumente in den Hintergrund rücken und eben die Dokumente als Grundlage verwenden, die einfach aufgefunden werden können. Die Chance für ein Konkurrenzangebot liegt hier in der Betonung der Qualität der erschlossenen Informationen.

Auch hinsichtlich der Dateiformate beschränkt sich Google Scholar nicht auf die für Texte üblichen Formate, sondern nimmt beispielsweise auch Powerpoint-Präsentationen mit in das Angebot auf. Auch hier ist natürlich keine Qualitätsbeurteilung gegeben; wie auch sonst durch den vielfachen Gebrauch von Powerpoint ergeben sich Tendenzen zur Übernahme von Schlagworten und aus dem Zusammenhang gerissenen Sätzen.

Hinsichtlich der Herkunft der Inhalte hat Google Scholar keine selbst gesetzte Grenze. Der sprachliche Schwerpunkt liegt – bei wissenschaftlichen Inhalten nicht verwunderlich – bei englischen Dokumenten. Deutschsprachige Dokumente sind zwar auch zu finden, aber in weit geringerem Maß. Da Google Scholar generell eine Tendenz zu haben scheint, die harten Wissenschaften, in denen die Publikation in englischer Sprache eher die Regel ist, zu bevorzugen (s.o.), kann nicht ermittelt werden, ob nur die Tendenz besteht, bestimmte Wissenschaften nur eingeschränkt zu berücksichtigen oder aber die Geistes- und Gesellschaftswissenschaften gerade aufgrund der vielfältigen Sprachen geringere Berücksichtigung finden.

Über den Umfang des Datenbestands werden von Seiten der Firma Google keine Angaben gemacht. Hochrechnungen aufgrund von Testanfragen besitzen nur eingeschränkte Gültigkeit, da die Trefferzählungen des Google-Systems als generell unzuverlässig gelten und die genannten Trefferzahlen oft nicht konsistent sind. So lässt sich leider nur eine grobe Schätzung über die Größe des Datenbestands abgeben: Diese liegt zwischen zwei und sieben Millionen Dokumenten, ist also im Vergleich zum Web-Bestand Googles oder auch dem geschätzten Gesamtumfang proprietärer wissenschaftlicher Quellen als gering anzusehen. Für die Zukunft ist allerdings eine wesentliche Erweiterung zu erwarten, einerseits durch die Zugewinnung neuer Verlagspartner, andererseits durch ein intensiveres Crawling wissenschaftlicher Quellen. Insbesondere die Einbindung von Open-Access-Archiven dürfte die Zahl der Dokumente noch einmal wesentlich erhöhen.

Abbildung 1 zeigt die Verteilung der Dokumente in Google Scholar nach Jahren. Ein Schwerpunkt liegt klar auf Dokumenten der letzten Jahre, insbesondere ist hier ein hoher Anteil frei zugängliche Quellen zu verzeichnen. Je weiter man auf der Zeitleiste zurückgeht, desto mehr Dokumente sind nur noch als aus den Literaturlisten anderer Dokumente extrahierte Literaturangaben vorhanden und können nicht im Volltext eingesehen werden. Die Verteilung erscheint nicht ungewöhnlich und spiegelt auch die stetig zunehmende Zahl wissenschaftlicher Veröffentlichungen im Lauf der Jahre wieder.

Wie oft der Datenbestand aktualisiert wird, konnte aufgrund des kurzen Zeitraums zwischen dem Start von Google Scholar und der Erstellung dieser Expertise nicht ermittelt werden. Allerdings liegt keine kontinuierliche Aktualisierung (wie etwa beim Web-Bestand von Google) vor; zumindest seit dem 1. Februar fand bis zum Abschluss dieser Arbeit keine Aktualisierung statt.

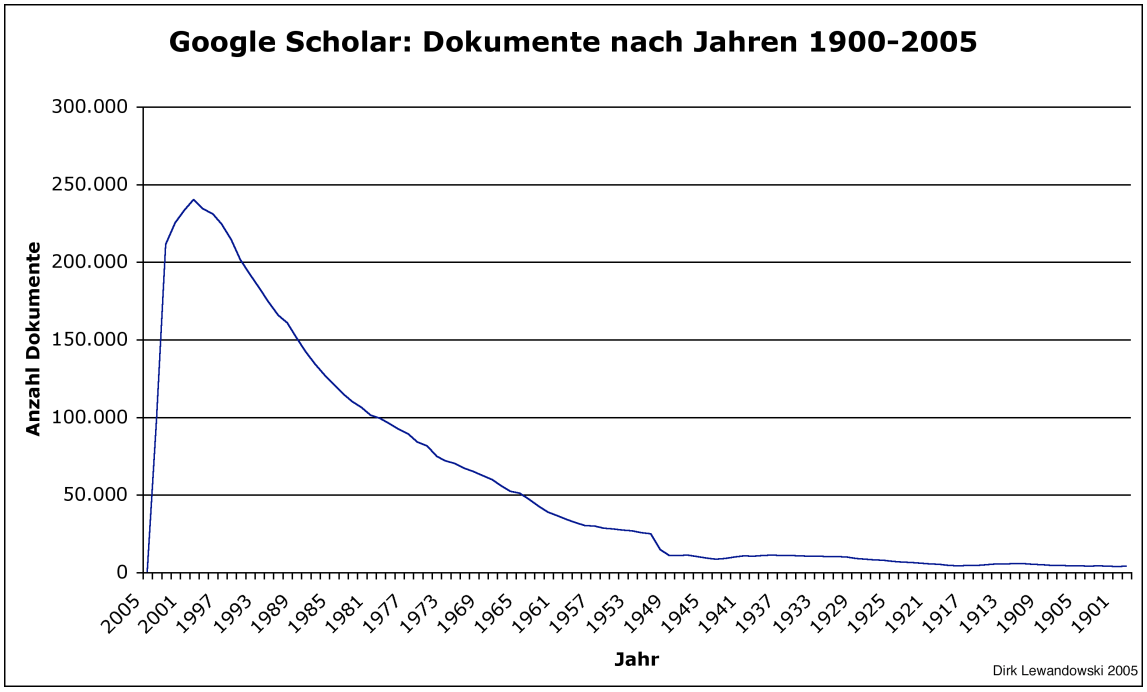

Abb. 1: Verteilung der Dokumente in Google Scholar nach Jahren (1900-2005)

Ein großes Problem für alle Suchmaschinen stellen Dubletten oder Fast-Dubletten dar. Viele wissenschaftliche Arbeiten werden an unterschiedlichen Stellen des Web in teils unterschiedlichen Versionen veröffentlicht. Zwar sollen solche Dubletten von Google Scholar erkannt werden, in der Praxis tauchen Dubletten allerdings in nicht geringer Zahl in den Trefferlisten auf. Solche Treffer sind für die Nutzer bei der Durchsicht der Trefferlisten lästig – ein alternatives System mit einer zuverlässigen Dublettenkontrolle würde hier ein besseres Bild hinterlassen.

3.2. Einbindung der Inhalte

Alle in Google Scholar verfügbaren Inhalte wurden durch Crawler erfasst. Das heißt, dass in dem Augenblick, in dem eine Suchanfrage gestellt wird, nur eine einzige Datenbank abgefragt wird, nämlich die, in der alle Informationen über die der Suchmaschine bekannten Dokumente abgelegt sind. Um diese Art der Erfassung zu ermöglichen, müssen die Crawler Zugang zu allen Dokumenten haben, die in den Bestand aufgenommen werden sollen. Im Fall der Inhalte von Verlagen und teils auch von Fachgesellschaften, die ihre Inhalte nur kostenpflichtig anbieten, sind deshalb Vereinbarungen nötig, die es zulassen, dass die Crawler der Suchmaschine (im Gegensatz zu menschlichen Nutzern) Zugang in geschützte Bereiche der jeweiligen Angebote erhalten.

Die Crawling-Lösung bringt den Vorteil mit sich, dass der Nutzer bei der Suche keinen Unterschied zwischen den unterschiedlichen Datenlieferanten bemerkt und dass die Performanz der Suchmaschine nicht von den Datenlieferanten abhängig ist, da zum

Zeitpunkt der Suchanfrage ja nur die Ressourcen der Suchmaschine selbst abgefragt werden.

Die Crawling-Methode erleichtert ebenso die Zusammenstellung und die Aktualisierung von Datenbanken, sofern eben für die Crawler des eigenen Projekts Zugang zu den entsprechenden Daten besteht. Dieser Ansatz ist dem der Metasuche vorzuziehen, da die Suche für den Nutzer wie bei anderen Suchmaschinen gewohnt und damit ohne lange Wartezeiten abläuft.

3.3. Kooperation mit Verlagen und Fachgesellschaften

Wie im letzten Abschnitt bereits dargestellt wurde, ist es für Google Scholar möglich, Inhalte zu spidern, die normalen Internet-Nutzern nicht oder nur eingeschränkt zugänglich sind. Dazu sind Partnerschaften mit Verlagen nötig. Hierbei kann Google auf seine Erfahrungen aus den Verlagspartnerschaften aus dem Google-Print- als auch aus dem CrossRef-Projekt nutzen.

Über die Anzahl und die Namen der Partner werden von Seiten Googles keine Angaben gemacht. Aus Tests ergibt sich aber, dass von den größeren wissenschaftlichen Verlagen und Gesellschaften zumindest ACM, Blackwell, Institute of Physics, Nature Publishing Group, Wiley Interscience, Springer und IEEE dabei sind. Auf der anderen Seite sind bspw. die Titel von Elsevier nicht enthalten. Es kann angenommen werden, dass Elsevier seine Bereitschaft versagte, um seine eigenen Angebote nicht zu gefährden.

Auch bei den aufgeführten Verlagen und Fachgesellschaften ist nicht klar, ob deren Inhalte vollständig oder nur zum Teil in Google Scholar erfasst werden. Wie in anderen Fällen auch ist auch hier Google mangelnde Transparenz vorzuwerfen.

3.4. Erschließung der Dokumente

Google Scholar erfasst alle Dokumente im Volltext. Damit werden wesentlich mehr Begriffe suchbar als dies bei einer Erschließung nur mittels formaler Sprachen möglich ist. Andererseits ist dadurch eine Relevanzbeurteilung der Dokumente nötig, da die Menge der Treffer entsprechend steigt und in einer beispielsweise nach Aktualität sortierten Trefferliste Dokumente von sehr unterschiedlicher Relevanz nebeneinander stehen würden. Die Relevanzbewertung der Dokumente basiert auf der Standard-Technik von Google, welche für den vorgegebenen Zweck modifiziert wurde s. Kap. 2). Die Anordnung der Trefferliste lässt sich nicht manuell ändern.

Die Volltexterschließung mit Relevanzbewertung bietet den Vorteil, dass einfach zumindest eine gewisse Menge an Dokumenten gefunden werden kann. Wenn ein Nutzer „mal eben“ ein paar Dokumente zu einem Thema benötigt, um sich beispielsweise einen Überblick zu verschaffen, ist ein solches System gut geeignet. Es ist intuitiv zu bedienen und es sind keine Kenntnisse über Klassifikationen, Thesauri oder andere Formen eines kontrollierten Vokabulars nötig.

Der Nachteil der alleinigen Volltexterschließung liegt in den geringen Möglichkeiten einer gezielten Suche. Mächtige Erschließungsinstrumente wie Klassifikation und Thesauri bieten gezielte thematische Einstiege innerhalb eines Fachgebiets und erlauben eine punktgenaue Suche. Viele der in Google Scholar erschlossenen Dokumente bieten in ihrer ursprünglichen Version (auf den Servern der Verlage bzw. Fachgesellschaften) in den Metadaten Klassenzuordnungen oder ergänzende Schlagworte. Diese Informationen werden von Google Scholar nicht berückscihtigt. Damit gehen diese für die Suche wertvollen Informationen für die Recherche verloren. Einem Konkurrenzsystem sollte die Synthese von intuitiver Suche und mächtiger Erschließung gelingen.

Bei der Aufbereitung der Volltexte der Dokumente geht Google den gleichen Weg wie bei seiner regulären Web-Suche: es erfolgt keine weitere Bearbeitung der Dokumente etwa durch Stemmingverfahren. So ergibt beispielsweise eine Suche nach „search engine“ und „search engines“ sowohl unterschiedliche Trefferzahlen als auch ein unterschiedliches Ranking.

Die Ausnahme von der reinen Volltexterschließung ist die Extrahierung der Autoren- und Zeitschriftennamen. Die Autoren werden in der Form Nachname Initialen angesetzt, bei den Zeitschriften wird die gebräuchliche Abkürzung (oder besser: *eine* der gebräuchlichen Abkürzungen) verwendet. Während diese Daten von den Partnerangeboten leicht und zuverlässig übernommen werden können, tauchen bei der Übernahme aus dem freien Web Probleme auf. Viele Dokumente werden falsch zugeordnet, die häufigsten Fehler sind die Vertauschung des Vor- und Nachnamens des Autors, die Verbindung von Nachname und Vorname von zweit Autoren zu einem neuen Namen und die Fehlerkennung des Autorenfelds, was zu Namen wie zum Beispiel „B Informationswirtschaft“ führt. Solche Fehlzuordnungen sind relativ häufig; für das genannte Beispiel werden 29 Dokumente gefunden.

4. Suchfunktionen

Wie bereits dargestellt, orientiert sich Google Scholar in seinem Aufbau und seinen Möglichkeiten aqn dem Hauptangebot von Google. Ebenso verhält es sich bei den Suchfunktionen und deren Strukturierung. Neben der einfachen Suche auf der Startseite, die nur ein Suchfeld umfasst, wird eine erweiterte Suchmaske angeboten, mit der genauere Suchanfragen durch die Kombination verschiedener Felder gestellt werden können. Daneben ist es möglich, mittels einer Abfragesprache erweiterte Suchanfragen auch in der Standardsuche zu formulieren. Dabei treten ähnliche Einschränkungen auf wie auch in der Google-Web-Suche; beispielsweise werden Boolesche Operatoren nur eingeschränkt unterstützt und die Feldbeschränkungen sind nur eingeschränkt kombinierbar.

Google Scholar bietet neben der Möglichkeit, einfach Suchbegriffe aneinanderzureihen (was eine automatische Verbindung dieser Begriffe mit UND zur Folge hat) auch die Möglichkeit, Begriffe auszuschließen (mit -), nach Phrasen zu suchen (indem die Begriffe in Anführungszeichen gestellt werden) und die Verknüpfung mit ODER. Diese ist allerdings dahingehend eingeschränkt, dass sich nur maximal zwei Begriffe mit ODER verbinden lassen, eine längere Verkettung ist nicht möglich. Auch die Klammersetzung zur Formulierung von Booleschen Argumenten wird nur eingeschränkt unterstützt. Vor allem auch im erweiterten Suchformular sind der Formulierung komplexerer Suchanfragen enge Grenzen gesetzt.

Die feldbeschränkte Suche hängt natürlich stark von den überhaupt erschlossenen Feldern ab (s. 3.4). Es ist möglich, die Suche auf den Titel, den Autor, die Publikation und das Jahr einzuschränken.

Titel: Die Titelsuche kann über das erweiterte Suchformular sowie über den Befehl *allintitle:* ausgeführt werden. Die Zuordnung der Titel erfolgt zuverlässig. Weitere (naheliegende) Beschränkungen wie die Beschränkung auf das Vorkommen der Suchwörter im Abstract bestehen nicht. Überhaupt erscheint es in einem wissenschaftlichen Kontext sonderbar, dass in Google Scholar die bei einem Großteil der Dokumente vorhandenen Abstracts vollkommen unberücksichtigt bleiben, sei dies bei der Suche oder in der Trefferlistenanzeige.

Autor: Auch diese Suchfunktion kann über das erweiterte Suchformular oder über einen eigenen Befehl (*author:*) ausgeführt werden. Diese Suche ist allerdings aufgrund der Menge fehlerhafter Zuordnungen nicht zu empfehlen.

Publikation: Diese Suche ist über das erweiterte Suchformular möglich; danach erscheint oberhalb der Trefferliste neben dem Feld für die reguläre Textsuche ein weiteres, in dem der Name der Publikation eingegeben werden kann. Auch die Zuordnung der Publikationen ist so fehlerbehaftet, dass eine Suche nach Artikeln, die in einer bestimmten Zeitschrift erschienen sind, nicht empfohlen werden kann. Ein Beispiel: Es soll nach Artikeln aus der Zeitschrift „Journal of the American Society for Information Science and Technology“ (JASIST) gesucht werden. Die Suchanfrage JASIST ergibt 13 Treffer, die Anfrage mit dem vollständigen Zeitschriftentitel ist nicht möglich, da das Suchfeld nicht genügend Zeichen zulässt. Zwar kann mit einem Teil

des Titel gesucht werden, dabei tauchen jedoch auch alle Aufsätze aus JASIS (dem Vorgänger von JASIST, ohne „Technology“) auf, die vielleicht in der Suche gar nicht gewünscht waren. Sucht man nach Artikeln aus JASIS, so erhält man 57 Treffer, gibt man den vollständigen Titel ein, sind es ca. 2.200.

Datum: Die Datumsbeschränkung erfolgt nach Jahren und über das erweiterte Suchformular. Auch hier werden oberhalb der Trefferliste nach dem Abschicken einer entsprechenden Anfrage die für die Datumsbeschränkung notwendigen Felder dargestellt. Die Beschränkung kann auf ein einziges Jahr oder auf einen Zeitraum erfolgen. Zwar liefert eine solche Beschränkung durchaus sinnvolle Ergebnisse, es kommt aber auch zu vielen Fehlzuordnungen. Dies ist insbesondere erstaunlich, als dass diese auch bei Informationen auftreten, die von Anbietern kommen, deren Angebote in großem Umfang indexiert werden und bei denen das Publikationsdatum immer an der gleichen Stelle steht und mit der gleichen Phrase eingeleitet wird. Als Beispiel sei hier das Portal der ACM genannt. Alle Jahresangaben werden mit „Year of Publication:“ eingeleitet, trotzdem kommt es zu falschen Zuordnungen, beispielsweise wird die Bandangabe als Jahreszahl übernommen.

Neben der Suche wird von Google Scholar auch das Browsing entlang von Zitationen unterstützt. In den Trefferlisten wird zu jedem Treffer angegeben, wie oft dieser von anderen Dokumenten zitiert wird. Klickt man auf diese Angabe, so erhält man eine Liste aller dieser Dokumente. Diese Funktion kann hilfreich sein, um thematisch verwandte Dokumente aufzufinden oder Entwicklungslinien in der Forschung nachzugehen. Aufgrund der vielen Dubletten und der ungenauen Erfassung eignet sich diese Funktion allerdings nicht für informetrische Recherchen.

Abschließend noch einige Beispiele für gängige Sucheinschränkungen, die von Google Scholar *nicht* unterstützt werden:

- Sprache: Es ist nicht möglich, nur nach Dokumenten in einer bestimmten Sprache zu suchen.
- Thematische Suche: Auch die Suche nach Dokumenten zu einem Thema ist nicht möglich, allein die Suche nach in den Dokumenten vorkommenden Stichwörtern. Dies ist ein grundsätzliches Problem aller Suchmaschinen(anbieter); die thematische Suche wird offenbar schlicht übersehen oder in ihrer Bedeutung verkannt.
- Einschränkung nach Texttyp: Eine Einschränkung nur auf Bücher, Artikel oder Präsentationen ist nicht möglich. Dies bedeutet auch, dass sich keiner dieser Texttypen ausschließen lässt, was vor allem bei Präsentation zu wünschen wäre.
- Einschränkung nach Qualität: In der Suche kann nicht zwischen den unterschiedlichen Versionen der Dokumente unterschieden werden. Es ist nicht möglich, beispielsweise nur nach den endgültigen, in Zeitschriften oder Konferenzbänden veröffentlichten Artikeln zu suchen. Es kann nicht verhindert werden, dass solche Treffer mit studentischen Arbeiten o.ä. vermischt werden.

5. Bewertung des Systems und Handlungsempfehlungen für ein umfassendes wissenschaftliches Informationssystem

In den folgenden Abschnitten werden die in den vorangegangenen Kapiteln dargestellten Eigenschaften von Google Scholar in Hinblick auf die Erstellung eines Konkurrenzsystems bewertet und entsprechende Handlungsempfehlungen gegeben.

5.1. Geschwindigkeit

Google Scholar besticht durch seine hohe Geschwindigkeit, die Antwortzeiten auch bei komplexen Anfragen liegen deutlich unter einer Sekunde.

Empfehlungen für ein Konkurrenzsystem:

- **Geschwindigkeit**: Um eine hohe Geschwindigkeit zu erreichen, sollten die Repräsentationen der zur Verfügung stehenden Dokumente in einer Datenbank zusammengefasst werden, was die Suche entsprechend beschleunigt. Eine Metasuche über die integrierten Angebote hinweg führt zu langen Bearbeitungszeiten, zumal wenn auch eine Dubletteneliminierung stattfinden soll. Die Nutzer sind in der Regel nicht bereit, „lange“ (also länger als einige Sekunden) auf die Suchergebnisse zu warten.

5.2. Inhalte

Zwar verfügt Google Scholar über eine nicht geringe Anzahl von Dokumenten, allerdings sind weder der genaue Umfang, die genauen Quellen noch die Vollständigkeit der erschlossenen Quellen klar. Google Scholar stützt sich auf drei Quellentypen: freies Web, Angebote von Verlagen und Fachgesellschaften, Open-Access-Server.

Empfehlungen für ein Konkurrenzsystem:

- **Transparenz der erschlossenen Quellen**: Dem Nutzer muss ersichtlich sein, welche Inhalte im System überhaupt recherchiert werden können. Dies könnte über eine hinterlegte Quellenliste inkl. eventueller Angaben über die erschlossenen Jahrgänge realisiert werden. Ein Nutzer, der in seinen Recherchen Vollständigkeit anstrebt, sollte wissen, welche Quellen er mit einer Recherche im System wie vollständig abgedeckt hat.
- **Erweitertes Quellenspektrum**: Im Sinne eines Single Access Point sollte eine Recherche über unterschiedliche Quellentypen (wie Verlagsinhalte, Open-Access-Server, usw.) möglich sein. Insbesondere sollte ein vollständiger Zugang zu den von öffentlichen Einrichtungen vorgehaltenen Archiven ermöglicht werden. Die Möglichkeit der Einbindung von Inhalten aus dem freien Web ist zu prüfen.
- **Abdeckung des gesamten Fächerspektrums**: Eine Antwort auf die Tendenz von Google Scholar zu technischen und naturwissenschaftlichen Inhalten in englischer Sprache sollte eine möglichst vollständige Abdeckung des Fächerspektrums sein –

betont werden sollten auch die Gesellschafts- und Geisteswissenschaften, die bevorzugt in der jeweiligen Landessprache publizieren.

5.3. Erschließung

Google Scholar stützt sich bei der Erschließung der Dokumente nahezu ausschließlich auf die Volltexterschließung; eine gezielte Suche ist kaum möglich. Die gefundenen Treffer werden nach Relevanz sortiert angezeigt.

Empfehlungen für ein Konkurrenzsystem:

- **Ranking**: Die Nutzer sind von allen Web-Suchmaschinen Ranking-Mechanismen gewöhnt, die auch bei der Eingabe nur sehr ungenauer Suchanfragen zumindest einige relevante Treffer liefern. Ein entsprechendes Rankingverfahren sollte eingesetzt werden, allerdings nicht als alleinige Möglichkeit. Daneben sollte es möglich sein, die Trefferliste nach eigenen Wünschen zu sortieren, vor allem die Sortierung nach dem Datum sollte möglich sein (und prominent platziert werden).
- **Tiefe Indexierung**: Neben dem Volltext sollten auch weitere Angaben erschlossen werden. Alle Dokumente sollten die (bereits von den Anbietern erstellten) Systemstellen, Deskriptoren, usw. behalten, um einerseits die zielgerichtete Recherche zu erleichtern, andererseits um diese im Ranking berücksichtigen zu können. Die nahezu alleinige Erschließung des Volltexts bei Google Scholar bietet für ein Konkurrenzsystem die Chance, durch eine verbesserte Erschließung den Nutzern qualitativ überlegene Treffer zu liefern. Dabei muss allerdings darauf geachtet werden, dass die intuitive Bedienbarkeit des Systems nicht darunter leidet.

5.4. Nutzerführung

Google Scholar verwendet eine schlichte Einstiegsseite mit nur einem Eingabefeld; erweiterte Suchfunktionen können über die Abfragesprache oder ein erweitertes Suchformular angesteuert werden. Das System ist klar auf die Suchfunktionalität ausgelegt, Browsing-Komponenten finden sich nur durch die Möglichkeit, sich vorwärts entlang der Zitationen zu bewegen.

Empfehlungen für ein Konkurrenzsystem:

- **Suchformulare**: Die Suchformulare sollten sich an der Einfachheit von Google Scholar orientieren; auf die Interface-Gestaltung (und Nutzerführung allgemein) sollte höchster Wert gelegt werden. Um sowohl die Bedürfnisse der ungeübten Nutzer als auch die der Profis zu berücksichtigen, sollte es neben dem erweiterten Suchformular auch möglich sein, Anfragen direkt in einer Abfragesprache einzugeben. Alle Suchfunktionen sollten beliebig kombinierbar sein.
- **Suche und Browsing**: Neben den Suchfunktionalitäten sollten auch Browsing-Komponenten integriert werden. Damit wäre es beispielsweise möglich, innerhalb von Klassifikationen, Thesauri, Publikationen und einzelnen Datenbanken zu navigieren.

- **Kostentransparenz**: Der Nutzer sollte stets im Bilde sein, welche Inhalte er kostenlos abrufen kann (sei es, dass diese generell kostenfrei angeboten werden oder dass er freien Zugang über seine Institution hat) und für welche er bezahlen muss. Diese Informationen sollten schon in der Trefferliste ersichtlich sein.
- **Verfügbarkeitstransparenz**: Auch die Verfügbarkeit (sofort digital vs. Bestellung und spätere Lieferung) sollte einfach ersichtlich sein.

5.5. Suchfunktionen

Google Scholar bietet nur geringe Suchmöglichkeiten, die teils auch noch mit Fehlern behaftet sind. Der geringe Umfang der Suchmöglichkeiten wird durch das Rankingverfahren teils kompensiert.

Empfehlungen für ein Konkurrenzsystem:

- **Suchmöglichkeiten**: Das System sollte möglichst umfassende Suchmöglichkeiten bieten, ohne den Nutzer zu einer elaborierten Suche zu zwingen. Auch die Eingabe nur weniger Suchbegriffe sollte zu brauchbaren Ergebnisse führen. Neben den üblichen Einschränkungsmöglichkeiten nach Autor, Systemstelle, Publikation, usw. sollte auch die Möglichkeit gegeben sein, nur qualitätsgesicherte Inhalte zu suchen (also in der Regel solche, die ein Peer-Review-Verfahren durchlaufen haben).

5.6. Mehrwerte

Google Scholar bietet zur Suche und der Weiterleitung zum Volltext keine Mehrwerte. Nutzer können an ein System gebunden werden, wenn es ihnen einen Mehrwert bietet, der die tägliche Arbeit erleichtert. Solche Mehrwerte können (bei ansonsten vergleichbarer Qualität von unterschiedlichen Systemen) über die bevorzugte Nutzung eines der Systeme entscheiden.

Empfehlungen für ein Konkurrenzsystem:

- **Alert-Service**: Neben der Suche durch den Nutzer (Pull-Ansatz) sollte die Einrichtung von Alerts (Push-Ansatz) möglich sein. Der Nutzer speichert seine Suche ab und erhält in regelmäßigen Intervallen die bibliographischen Angaben (und evtl. Abstracts) der neu ins System eingestellten Dokumente zu seinen Interessengebieten per E-Mail zugeschickt. Der Nutzer erhält damit nicht nur einen Mehrwert, sondern wird auch mit jeder E-Mail an die Nutzung des Systems erinnert. Google Scholar bietet bisher keinen Alert-Service, auf Dauer ist jedoch mit einem solchen zu rechnen, da dieser bereits für die Websuche und für Google News verfügbar ist.
- **Weitere Mehrwertdienste**: Eine weitere Maßnahme der Nutzerbindung wäre die Möglichkeit, aus den Trefferlisten direkt Bibliographielisten zur Einbindung in eigene Arbeiten zu generieren. Einen solchen Service bieten unterschiedliche kommerzielle Datenbank-Anbieter bereits an, z.B. Cambridge Scientific Abstracts (CSA). Ein solches Angebot wird von Google Scholar nicht gemacht und könnte Nutzer dazu bewegen, das System zu wechseln. Auch dürfte sich eine solche

Funktion, die der Bequemlichkeit der Nutzer entgegenkommt, schnell im Kollegenkreis herumsprechen.

6. Abschließende Bemerkungen

Google kann sich bei der Einführung von Google Scholar auf bereits etablierte Elemente des Google-Systems stützen. Die Nutzer sind mit der Bedienung des Systems vertraut und brauchen sich weder in die Nutzung von Google Scholar einzuarbeiten noch sich an ein neues System gewöhnen.

Die Berücksichtigung des durch die Nutzung von Suchmaschinen generell veränderten Nutzerverhaltens ist als der Schlüssel für den Erfolg von neu gestalteten Informationssystemen anzusehen. Dass die Nutzer bei ihren Recherchen mittlerweile weniger genau vorgehen als früher und erwarten, dass ihnen innerhalb von Sekunden Ergebnisse im Volltext angezeigt werden, mag man bedauerlich oder vermessen finden, letztlich müssen Informationssysteme allerdings auf die Nutzer ausgerichtet werden. Dazu gehört es, sich von dem Paradigma der Vollständigkeit der Ergebnisse bei allen Suchen zu verabschieden. Vielmehr sollten einerseits die genannten „quick and dirty"-Recherchen unterstützt werden, bei denen der Nutzer „mal eben schnell" einige Dokumente zu seinem Thema finden möchte. Andererseits sollten auch komplexe Recherchen unterstützt werden und damit Möglichkeiten gegeben werden, die Systeme wie Google Scholar nicht bieten.

Die Zeit bis zur offiziellen Einführung von Google Scholar und dem damit verbundenen Popularitätsschub sollte genutzt werden, um ein Konkurrenzangebot so zu optimieren, dass es zu diesem Zeitpunkt mit dem Google-Angebot in allen der oben als bedeutend dargestellten Punkten konkurrieren kann. Zum Startzeitpunkt werden die Nutzer Google Scholar mit den ihnen sonst bekannten Systemen vergleichen bzw. bei der erstmaligen Nutzung zumindest die Möglichkeiten des bisher von ihnen verwendeten Systems vor Augen haben. Diese Tests werden zugunsten der von den Nutzern als ihren Bedürfnissen gemäß geeigneteren Suchmaschine ausfallen.