Data Smog, Precision und Recall: Retrievalstrategien zur Ballast-Reduzierung bei Internet-Recherchen

Autor: Dr. Christopher N. Carlson
IWF Wissen und Medien gGmbH
Telephon: +49-551-5024-311
Fax: +49-551-5024-322
E-Mail: christopher.carlson@iwf.de

Abstract:

Die sog. Informationsüberflutung - auch bisweilen *data smog* genannt - erweist sich zunehmend als Problem beim Internet-Retrieval. Technikbasierte Lösungsansätze wie personalisierte Agenten, Filter, Ranking-Algorithmen oder infometrische Analysetechniken haben zwar eine anfängliche Euphorie ausgelöst, letztlich jedoch das Problem nicht lösen können. Bei der derzeit zu beobachtenden Expansionsrate der Internet-Seiten nimmt das Signal-zu-Rausch-Verhältnis exponentiell ab. Selbst eine gezielte Reduzierung von Recall-Werten als Retrieval-Strategie verursacht in der Regel Ballast-Mengen, die eine qualifizierte Relevanzprüfung kaum noch zulassen.

Auch greifen Forderungen nach mehr Standardisierung von Datenstrukturen in Webinhnalten oder nach konsequenter Anwendung von Metatag-Funktionen zu kurz, da sie unter den reellen Produktionsbedingungen des WWW nicht praktikabel sind. Der Vortrag will alternative Strategie-Ansätze präsentieren, ihre Stärken und Schwächen darlegen, sowie aufzeigen, für welche Fragestellungen das Web jetzt und in absehbarer Zukunft ein geeignetes Suchobjekt sein kann. Insbesondere werden syntagmatische und erfahrungsbasierte Retrieval-Strategien demonstriert

---

1. *Data Smog* - ein Problem?

Informationsüberflutung ist ein Phänomen mit sowohl objektiven als auch subjektiven Ursachen. Objektiv gesehen, ist die Menge leicht erhältlicher Information im Verlauf der letzten fünfzig Jahren in jedem einzelnen Jahrzehnt exponentiell angewachsen. Derzeit gibt es keine Anhaltspunkte dafür, daß sich diese Wachstumsraten in absehbarer Zukunft verringern werden. Dagegen ist grundsätzlich auch nichts einzuwenden, handelt es sich doch um das normal zu erwartende Ergebnis einer Kombination aus einem freien Informationsmarkt und IKT-Fortschritten.

Die subjektive Komponente der Informationsüberflutung kommt dadurch zustande, daß wir mehr Information zur Verfügung haben, als wir ohne weiteres verarbeiten können. Es handelt sich also um ein Rezeptionsproblem, das bisweilen als "Technostreß" bezeichnet wird.[1] Diese subjektive Wahrnehmung verursacht eine weitere, damit einhergehende Rezeption, derzufolge EDV-Benutzer durch ihre IKT beherrscht werden, anstatt daß sie das Gefühl haben, dadurch mehr Möglichkeiten zu bekommen. Wie mit jeder anderen Art von Streß auch, hat Technostreß eine reduzierte intellektuelle Leistung und ein herabgesetztes Urteilsvermögen zur Folge; eine Tatsache, die in der kognitiven Psychologie gut bekannt und beschrieben ist. In einer Art

[1] Rosen, Larry; Weil, Michelle: "Technostress: Coping with Technology @ Work, @ Home, @ Play". John Wiley & Sons, 1997.

negative Rückkoppelungsschleife ist dies teilweise Ursache, teilweise Resultat einer inkonsistenten und wenig zielführenden Nutzung von IKT. Das Nicht-Vorhandensein eines kohärenten Wissensmanagement-Rahmens mag erschwerend hinzukommen.

Hier einige Zahlen, die ein Schlaglicht auf die schier immensen Dimensionen werfen, um die es geht[2]:

3,062 - die Anzahl von Zeitungs- und Zeitschriftenbeiträgen aus den USA zwischen 1997 und 1999 die die Informationsüberflutung thematisieren

15,652 - die Anzahl von WWW-Fundstellen, die Informationsüberflutung behandeln

2,892 - die Anzahl von Titeln im Katalog der Library of Congress in denen das Wort *stress* vorkommt

454 - die Anzahl an Dokumenten, die pro Minute dem Internet-Dienst Lexis-Nexis hinzugefügt werden

40% - Prozentsatz der Berufstätigen, die von sich sagen, daß ihre Arbeit durch mehr als 6 ungewollte Mitteilungen pro Stunde unterbrochen wird

50% - Prozentsatz der Berufstätigen (in den USA), die häufig Nachrichten mit Wiederholungscharakter erhalten

190 - Anzahl der Nachrichten (unabhängig vom benutzten Nachrichtenweg), die der durchschnittliche Mitarbeiter einer Firma auf der *Fortune 1000*-Liste täglich verschickt oder empfängt

80% - Prozentsatz an Informationen, die abgelegt, aber nie benutzt werden

150 - Anzahl Stunden, die der Durchschnittsbürger jährlich mit der Suche nach verlorengegangenen Informationen verbringt

71% - Prozentsatz der Berufstätigen, die ihre Hauptarbeit als das Aufspüren von Informationen bezeichnen

44% - Prozentsatz der Führungskräfte, die der Ansicht sind, die Kosten der Informationsbeschaffung lägen höher als der Informationswert für ihr Geschäft

8 : 1 - Verhältnis der online verfügbaren Beiträgen zu denen in Tageszeitungen

17 - Anzahl der Seiten, um die das WWW jede Sekunde wächst

7,349,000 - Erwartete Zunahme an URLs für die Zeit zwischen 1997 und 2002

18,300,000 - Anzahl neuer Faxgeräte in den USA seit 1987

[2] Quelle: "Data Data". Inc Magazine; January 1, 1999. URL: http://www.inc.com/magazine/19990101/715.html (akzessioniert am 3. Juni 2003)

2,809,000 - Zunahme (in Tonnen) der Papiermenge, die zwischen 1984 to 1998 in Büros verwendet wurde

Historisch betrachtet, ist ein Mehr an Information fast immer eine sehr gute Sache gewesen. Information ermöglichte die Verbreitung der Kultur und die Entwicklung von Kommerz und Technik; sie war auch eine der treibenden Kräfte hinter der großflächigen Etablierung von Demokratie und Menschenrechten. Es gab buchstäblich keine schädlichen Nebenwirkungen, wenn man mehr Information hatte. Informationsbesitz brachte nur Vorteile. Seltsamerweise wird die Informationsgesellschaft nach etwas benannt, das einst uneingeschränkt positiv rezipiert wurde, was aber mittlerweile zunehmend als Problem empfunden wird.

Das Signal-zu-Rausch-Verhältnis ist eine häufig verwendete Metapher für die Beschreibung von Informationsüberflutung. Der Ausdruck stammt ursprünglich aus der Tonträgerindustrie und wurde benutzt, um das Verhältnis von gewolltem Ton - z.B. Musik - zu ungewollten Geräuschen wie etwa das Knistern älterer Vinyl-Schallplatten. Im Kontext der Informationsgesellschaft wird der Terminus verwendet, um das Verhältnis von nützlicher Information, die gefunden wird, zu der Gesamtmenge an gefundener Information zu charakterisieren. Abgesehen vom Problem der absoluten Zunahme an Information, gibt es das zusätzliche Problem der relativen Abnahme der Relevanz bzw. Triftigkeit (Pertinenz) gefundener Dokumente. Die Leichtigkeit sowie die geringen Kosten des Online-Publizierens - nachdem man schon einen maschinenlesbaren Text und eine eigene Website hat, sind die *zusätzlichen* Kosten ziemlich geringfügig - haben zu einer vorhersehbaren Informationsflut geführt, von der vieles beim besten Willen nur als trivial oder irrelevant bezeichnet werden kann. Tatsächlich existieren viele Websites, die sich selbst als "Useless Knowledge Page" und dergleichen bezeichnen[3]. Solche Seiten liefern Informationen wie, daß das epische indische Gedicht "Mahabhrata" achtmal länger als die Ilias und die Odyssee zusammen ist.[4] Aber auch ohne den absichtlichen Versuch, Irrelevanz zu erreichen, ist es offenkundig, daß jede Suchfrage zu einem beliebigen Thema in einer Art Realkonkurrenz zu unvorstellbaren Mengen an Informationen steht, die zwar irgendeine andere Frage beantworten könnte, aber nicht die aktuell vorliegende.

Dies bedeutet, daß das Signal-zu-Rausch-Verhältnis empfindlich zurückgegangen ist - mit dramatischen Auswirkungen auf Precision und Recall.

Etwas sehr Ähnliches ist auch mit E-Mail passiert. Die ihr zugrundeliegende Technologie ist an sich sehr für eine unkontrollierbare Proliferation empfänglich. Wie auch mit dem Online-Publizieren gibt es bei E-Mail keine wirklichen Zusatzkosten, nachdem man erst einmal einen E-Mail-Provider und einen maschinenlesbaren Text hat. Es ist ebenso einfach und billig, eine Nachricht an Hunderte oder an Tausende zu versenden, wie an einen einzigen Empfänger. Man braucht lediglich die Adressen in einem Verteiler zu organiseren. Das Zusammenführen von - teilweise eingekauften - Verteilern zu personalisierten Versandlisten ermöglicht es, mit nur wenigen Tastenschlägen unerwünschte Nachrichten an Menschen zu verschicken, die der Absender überhaupt nicht kennt. Angesichts dessen war es wohl unvermeidbar, nur eine Frage der Zeit also, bis jemand das "Spamming" - den nicht angeforderten Massenversand kommerzieller bzw. werblicher E-Mails - erfand.

[3] Allein 29 wurden durch die Google-Suchmaschine nachgewiesen. (akzessioniert am 18. Juni 2003)
[4] http://www.coolquiz.com/trivia/directory/directory.asp?dir=Miscellaneous (akzessioniert am 18. Juni 2003)

Die IKT-Geschichte überliefert ein folgenschweres Datum:[5]: Am 12. April 1994 haben Laurence Canter und Martha Siegel, ein Anwaltsehepaar aus Arizona in den USA, Spamming auf einen neuen Gipfel des Mißbrauchs gehoben, indem sie ein unaufgefordertes werbliches Angebot, über eine sog. Green-Card-Lotterie an über 6000 Usenet Newsgroups verschickten. Davor waren Spam-E-Mails eine eher sporadische und halbherzige Angelegenheit, in der Regel auf athematische Nachrichten in einzelnen Newsgroups beschränkt.

Nicht einmal zehn Jahre später sind Spams ein gravierendes Problem im Hinblick auf die Informationsüberflutung geworden. Nach Aussage einer neueren Erhebung[6] ist die Spam-Menge allein in den letzten 18 Monaten um das Fünffache gestiegen, wodurch die elektronischen Briefkästen von Internet-Benutzern mit Milliarden von unerwünschten Werbenachrichten überfüllt werden. AOL blockiert 780 Millionen Spam-Nachrichten täglich, also 100 Millionen mehr als tatsächlich zugestellt werden.

2. Technologie löst nicht das Problem

Rezente IKTen haben ein eindrucksvolles Arsenal entwickelt, um die Probleme zu behandeln, die durch Informationsüberflutung entstehen. Die wichtigsten technikbasierten Retrieval-Hilfen können entweder in Internet-Suchmaschinen integriert oder auf die von deren Suchergebnissen gemachten Downloads angewandt werden.

Welche Hilfswerkzeuge gibt es und wie funktionieren sie?

- Intelligente Agenten - Agenten sind meist Werkzeuge für Informations-Retrieval. Einige Applikationen sind intelligente Retrieval-Schnittstellen, mediatisiertes Suchen und Brokering, Cluster-Analyse und Kategorisierung. Mithilfe einer agentenbasierten Herangehensweise können Retrieval-Systeme skalierbarer und flexibler werden, sie können leichter erweitert werden und mehr Interoperabilität erlangen. Dies erfolgt durch Agenten, die Datenströme optimal leiten, Suchfragen an Drittsystem "vermakeln" und Metadaten miteinander teilen.
- Ranking-Algorithmen - Mithilfe von diesen Algorithmen kalkulieren Suchmaschinen die Positionen der Ergebnisanzeige. Ranking-Algorithmen können sehr verschiedene Gesichtspunkte wie Domänennamen, spider-fähige Inhalte, Submissionspraktiken, HTML-Quellencodes und Verlinkungshäufigkeiten berücksichtigen. Die genaue Gestaltung von Suchmaschinen-Ranking-Methoden wird meist als Betriebsgeheimnis behandelt und auch häufig modifiziert, um Websites herauszufiltern, die versuchen, die Ergebnisanzeige zu manipulieren.
- Cluster-Analyse - Dies ist eine statistische Untersuchungsmethode, die versucht, die "natürlichen" Gruppierungen von Objekten auf der Grundlage von Information über ihre Attribute zu finden. Typischerweise sind dabei die Objekte die Variablen und die einzelnen Ausprägungen (Entitäten) sind die Attribute. Das Ergebnis einer Cluster-Analyse ist eine graphische Darstellung, die Dendrogramme enthält. Diese zeigen die Gruppierungen auf, die sich in den Daten herausgebildet haben.
- Web Mining / Data Mining - Data Mining ist die datenbasierte Aufdeckung und Modellierung von verborgenen Mustern, die sich in größeren Datenmengen befinden. Data Mining unterscheidet sich insoweit von den vorgenannten retrospektiven

[5] Schenk, David: "Data Smog: Surviving the Information Glut". 1997. URL: http://www.salemstate.edu/~tevans/overload.htm (akzessioniert am 16. Juni 2003)
[6] Vise, David A.: "AOL Joins Microsoft In a Reply to Spam" 21.02.2003. URL: http://www.washingtonpost.com/ac2/wp-dyn?pagename=article&node=&contentId=A38150-2003Feb20&notFound=true (akzessioniert am 19. Juni 2003)

Techniken, weil es die Herstellung von Modellen ermöglicht, die diese verborgenen Muster abbilden. Mit Data Mining ist es möglich, Muster zu entdecken und Modelle zu kreieren, ohne genau vorher wissen zu müssen, wonach man sucht. Die gebildeten Modelle können beschreibend oder auch vorausschauend sein. Man kann hypothetische Fragen an ein Data-Mining-Modell stellen, die bei einer herkömmlichen Datenbasis oder Data-Warehouse nicht möglich wären.

- Webgraph-Algorithmen - Diese basieren auf einem häufigen Internet-Phänomen: Für ein beliebiges Thema gibt es meist bestimmte "autoritative" Seiten, die auf das Thema fokussiert sind, und weitere Verweisseiten, die Links auf weiterführende Seiten, die für das Thema relevant sind, enthalten. Diese Beobachtung motivierte die Entwicklung neuer Suchalgorithmen: Mithilfe von speziellen Ableitungsregeln werden Autoritäts-Hierarchien ermittelt und graphisch dargestellt.
- Personalisierung, Empfehlungen und kollaborative Filter - Personalisierungs-Algorithmen verwenden benutzerspezifische Informationen, um maßgeschneiderte Informationsangebote zu kreieren. Bekannte Nutzerpräferenzen werden durch die Auswertung von Logfiles früherer Anwendersitzungen oder aus Angaben, die in Nutzerprofilen stehen - oder beides ermittelt. Empfehlungen und kollaborative Filter führen dieses Prinzip weiter, indem sie es auf die Präferenzen anderer Nutzer ausdehnen. Die Datenbasis "weiß", daß signifikante Mengen von Nutzern, die sich für "A" interessieren, sich ebenfalls für "B" interessieren. Dies kann dazu verwendet werden, um "B" mit einer einer höheren Priorität zu belegen, wenn es zusammen mit "A" vorkommt, als dies normalerweise aufgrund der genauen Formulierung des Sucharguments der Fall wäre. Ein bekanntes Beispiel für Empfehlungen und kollaborative Filter ist die Amazon-Website, wo man die Information bekommt, daß Personen, die das Buch, das man sich gerade anschaut, gekauft haben, ebenfalls bestimmte andere Bücher gekauft haben. Es gilt dabei die Annahme, daß ähnliche Nutzerpräferenzen in konsistenter Weise ähnlich bleiben.

Suchmaschinen unterstützen recht unterschiedliche Retrievalstrategien. Infolgedessen hängt die Wahl der geeigneten Suchmaschine für eine Frage von dessen genauem Gegenstand ab. Für einen aktuellen Vergleich der retrievalrelevanten Attribute gängiger Suchmaschinen sei hier auf die Search Engine Showdown[7] Website verwiesen.

Ebenfalls in diesem Kontext erwähnenswert ist eine Disziplin, die als Informations-Design oder als Informations-Architektur bezeichnet wird, und die entstanden ist, um der Herausforderung zu begegnen, Information für eine zielführende Kommunikation verwertbar zu machen. Diese Disziplin umfaßt solch unterschiedliche Themenfelder wie Betriebswirtschaft, Informatik, kognitive Psychologie, graphisches Design und Typografie sowie die technische Dokumentation. Informations-Architektur heißt, Daten im Hinblick auf die informatorischen bzw. auf die Management-Anforderungen von Organisationen oder Personengruppen zu strukturieren.[8]

Obwohl diese verschiedenen Techniken durchaus palliative oder abhilfeschaffende Wirkung haben, vermögen sie dennoch nicht, das Problem zu lösen.

[7] Notess, Greg: "Search Engine Showdown. The Users' Guide to Web Searching". URL: http://www.searchengineshowdown.com/ (akzessioniert am 16. Juni 2003)
[8] Für weitere Informationen zum Thema siehe: Victor, Stephen P.: "Instructional Applications of Information Architecture". URL: http://library.thinkquest.org/50123/info_arch.html (akzessioniert am 3. Juni 2003)

Auch Anti-Spam-Filter und andere technikbasierte Lösungsansätze für das Problem unerwünschter kommerzieller E-Mails haben bestenfalls einen geringen Wirkungsgrad erzielt. Das Problem wird zusätzlich verschärft, weil Spam-Versender die funktionalen Methodologien dieser Werkzeuge genau studieren, um sie besser umgehen zu können.

Es hat nur tangentiell mit dem Thema zu tun, aber man sollte sich auch nicht zu viel von einer vermehrten Standardisierung von WWW-Inhaltsstrukturen oder von einer konsistenteren Anwendung von Metatags versprechen, da dies - so wünschenswert sie auch sein mögen - oftmals nicht praktikabel oder sogar unmöglich sind unter den realen Produktionsbedingungen des WWW.

3. Was hilft wirklich?

Der erste und vielleicht auch wichtigste Schritt zur erfolgreichen Bewältigung der Informationsüberflutung besteht darin, daß Nutzer ein besseres Verständnis des Internets als Informationsressource entwickeln sollten. Realistische Erwartungen über die Verfügbarkeit von Informationen tragen dazu bei, Technostreß und die damit einhergehenden Erscheinungen Enttäuschung und Frustration zu reduzieren.

Die Frage der Verfügbarkeit von Informationen im Internet hängt mit realistischen Erwartungen zusammen darüber, welche Typen von Informationen zu finden sein dürften. Aufgrund der immensen Datenmengen, die online abrufbar sind, könnte man leicht der Auffassung sein, daß alles dort zu finden sein müßte, wonach man suchen möchte. Es ist jedoch klar nachvollziehbar, daß das nicht der Fall sein kann. Mögliche Gründe für das Nicht-Vorhandensein einer bestimmten Ressource oder Information sind u.a.[9]:

- Autoren und Verlage, die durch die Erschaffung und Verbreitung von Informationen Einnahmen erzielen wollen, werden in der Regel den Informationsmarkt nicht unterlaufen und dieselben Informationen kostenlos im Internet anbieten wollen.
- Die Aufbereitung und Verfügbarmachung von Informationen sind zeitaufwendig und kostspielig. Also mögen eventuell Einrichtungen, die ihre Informationsangebote sonst unentgeltlich anbieten würden, aus Kostengründen dies eben nicht mittels Internet tun.
- Die Nachfrage nach einem bestimmten Informationstyp bestimmt häufig dessen Verfügbarkeit. Viele Informationen, die verfügbar gemacht werden könnten, werden nicht angeboten, weil die Nachfrage danach nicht sehr groß ist. Z.B. sind rezente Wahlergebnisse sehr viel wahrscheinlicher im Internet erhältlich als solche aus den dreißiger Jahren.
- Textbasierte Informationen sind verhältnismäßig leicht mittels des Internets zu verbreiten. Numerische, graphische oder av-mediale Informationen sind schwieriger verfügbar zu machen, weil sie nicht direkt sondern nur über ihre Deskribierungen recherchierbar sind. Daher sind textbasierte Informationen viel häufiger als nicht-textbasierte erhältlich.
- Bestimmte Informationstypen sind grundsätzlich nicht erhältlich - und zwar unabhängig vom Verbreitungskanal. Z.B. die Frage, wieviele Kameras hat Robert Flaherty bei der Verfilmung von "Nanook of the North" verwendet?

In diesem Zusammenhang ist das Phänomen zu erwähnen, das manchmal als "Deep Web" - bisweilen auch als "das verborgene Web" - bezeichnet wird. Die Datenbasen von Suchmaschinen werden größtenteils durch "Spinnen", Crawler oder andere agentenähnliche

[9] Vgl.: Hinchcliffe, Lisa J.: "The Electronic Library: The World Wide Web". URL: http://alexia.lis.uiuc.edu/~janicke/ElecLib.html. Update vom 29. Mai 1997. (akzessioniert am 6. Juni 2003)

Programme rekrutiert. Natürlich können diese nur statische Seiten finden. In der Regel werden nur Seiten gefunden, die auf größeren Domain-Name-Servern stehen sowie solche, die mit anderen gefundenen Seiten verlinkt worden sind.

Die folgenden Feststellungen wurden in einem Grundsatzreferat eines Mitglieds der BrightPlanet-Gruppe veröffentlicht[10]:

- Öffentlich zugängliche dynamische Seiten im Deep Web machen mengenmäßig das 400-bis 550-fache des World Wide Webs aus.
- Das Deep Web umfaßt 7500 Terabyte an Information verglichen mit nur 19 Terabyte im WWW.
- Das Deep Web enthält beinahe 550 Milliarden Einzeldokumente verglichen mit 1 Milliarde im WWW.
- Über 200000 Deep-Web-Sites existieren.
- Sechzig der größten Deep-Web-Sites enthalten zusammen ca. 750 Terabyte an Informationen - und sind somit um ein Vielfaches größer als das offizielle Web.
- Deep-Web-Sites erhalten durchschnittlich 50% mehr Seitenaufrufe als normale Websites und sind auch stärker verlinkt als diese. Dennoch sind die meisten Deep-Web-Sites den meisten Internet-Nutzern nicht sehr gut bekannt.
- Das Deep Web wächst schneller als das Internet insgesamt.
- Hochwertige (geprüfte) Inhalte sind laut einschlägigen Qualitätskennzahlen über 1000mal häufiger im Deep Web als auf konventionellen Websites anzutreffen.
- Über die Hälfte der Deep-Web-Sites haben Ihre Inhalte in themenspezifischen Datenbanken.
- Über 95% der Deep-Web-Informationsangebote können unentgeltlich akzessioniert werden.

Erschwerend hinzu kommt die geringe Überschneidungsmenge in den Datenbasen der beliebtesten Suchmaschinen. Die neueste Überschneidungs-Untersuchung von Greg Notess[11] fand heraus, daß 50% der relevanten Seiten von nur einer von insgesamt zehn verwendeten Suchmaschinen gefunden wurden. Weitere 21% wurden von nur zwei Suchmaschinen gefunden. Dies unbeschadet der Tatsache, daß Notess ebenfalls ein beachtliches Datenbasis-Wachstum in den zwei Jahren vor seinem jüngsten Untersuchungs-Update vom März 2002 festgestellt hatte. Nebenbei bemerkt, ist dies als indirekter Indikator des Ausmasses des Internet-Seitenwachstums insgesamt gut geeignet, denn wenn die zehn beliebtesten Suchmaschinen alle erhebliches Datenbasis-Wachstum verzeichnet haben, indessen gleichzeitig die Datenbasis-Überschneidung stagniert oder sogar abnimmt, dann ist es offensichtlich der Fall, daß die tatsächliche Zahl potentiell auffindbarer Seiten im Web in noch größerem Umfang gewachsen sein muß.

Angesichts dessen, daß das "Oberflächen-Web" dramatisch kleiner ist als das Deep Web, während zugleich die durchschnittliche Suchmaschine nur einen geringen Teil des Webs insgesamt in ihrer Datenbasis hat, ist es folglich klar, daß selbst ein perfekt formuliertes Suchargument die meisten verfügbaren und relevanten Dokumente nicht finden wird, wenn der Nutzer seine Retrieval-Strategie auf eine einzige Suchmaschine abgestellt hat.

---

[10] Bergman, Michael K.: "The Deep Web: Surfacing Hidden Value". aus: The Journal of Electronic Publishing; Vol. 7, Issue 1; August 2001. URL: http://www.press.umich.edu/jep/07-01/bergman.html (akzessioniert am 4. Juni 2003)

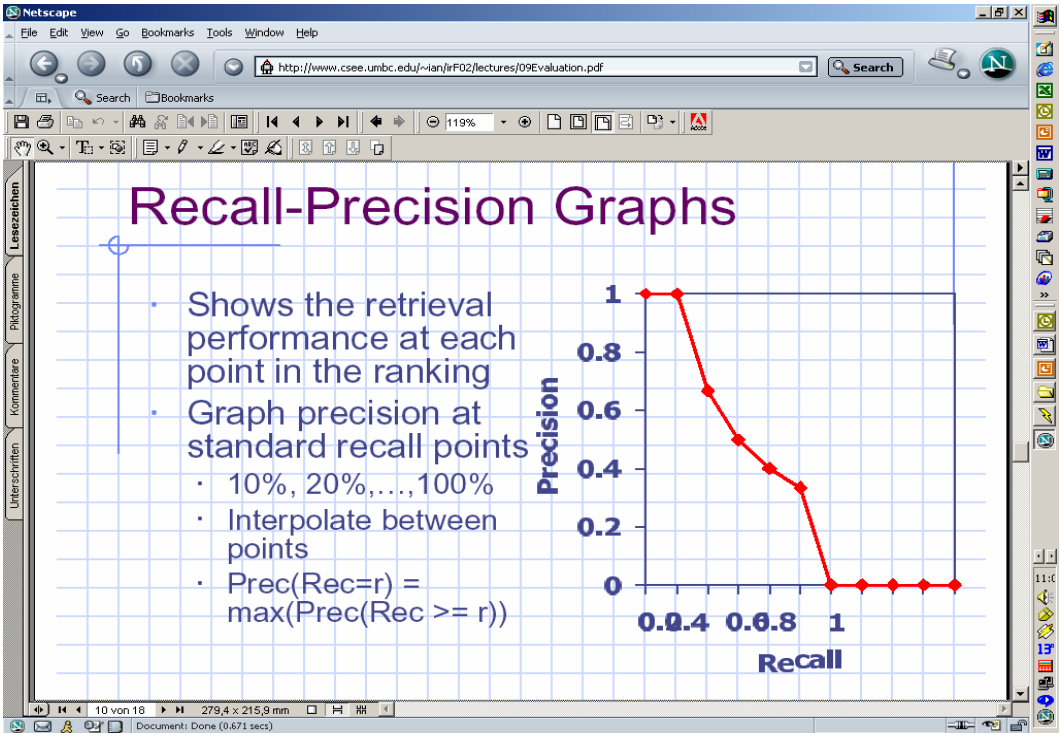
[11] G. Notess, a.a.O. URL: http://www.searchengineshowdown.com/stats/overlap.shtml (akzessioniert am 16. Juni 2003)

Obwohl es *prima vista* paradox erscheinen mag, wirkt sich das Nicht-Finden von relevanten Informationen, von denen man genau weiß (oder wenigstens erahnen kann), daß sie tatsächlich da sind, ebenso streßinduzierend aus wie der weitaus üblichere Überflutungs-Effekt einer unerwünschten Ballast-Menge.

Es ist daher von großer Wichtigkeit zu wissen, wie Suchmaschinen im allgemeinen funktionieren und woher ihre Datenbasen kommen. Da unterschiedliche Suchmaschinen auch unterschiedliche Werkzeuge benutzen, um ihre Retrieval-Ergebnisse zu optimieren, oder bestimmte Retrieval-Strategien eher unterstützen als andere, sollte man auch wissen, wie einzelne Suchmaschinen arbeiten. Obwohl das Phänomen der "Lieblings-Suchmaschine" recht weit verbreitet ist, ist es eindeutig nicht so, daß eine bestimmte Suchmaschine für alle - oder auch nur für die meisten - Suchprofile die beste ist. Die Benutzerkompetenz steigt, wenn man die Stärken und Schwächen von mehreren Suchmaschinen - darunter auch möglichst eine Meta-Suchmaschine - kennt, und diese auch wirksam einsetzen kann.

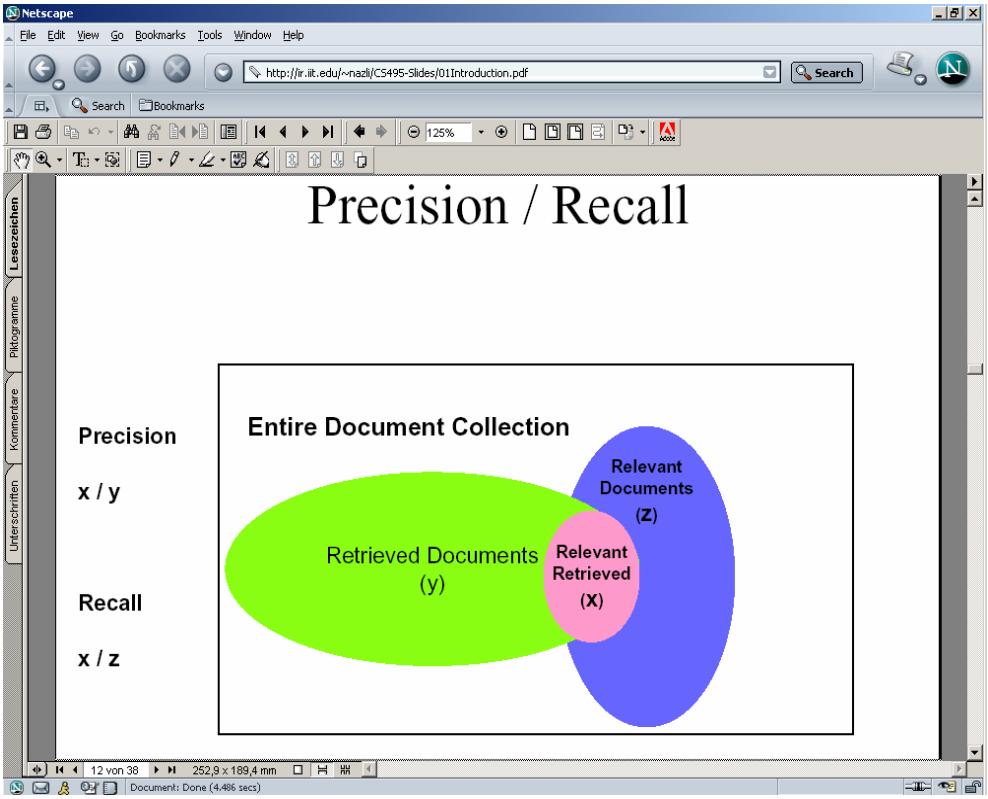
Was sagen uns Suchmaschinen-Ergebnisse über Precision und Recall aus? Welche alternativen Retrieval-Strategien legen die Ergebnisse eines ersten Sucharguments nahe?

Precision vs. Recall ist das Hauptparadigma für die Evaluation von Qualität im Informations-Retrieval. Idealerweise möchte man natürlich bei jeder Suchanfrage 100% Precision und 100% Recall haben. Das Dilemma der Informationsüberflutung ist aber, daß bei vielen Abfragen selbst ein relativ niedriger Recall-Wert fast unumgänglich zu einem Precision-Wert nahebei null führt. Z.B. wird ein einziges relevantes Dokument, das an sich R = 100% darstellt, faktisch in einer Treffermenge von ca. 1000 vollkommen untergehen (wobei die restlichen 999 falsch-positive Treffer sind), wenn es nicht gerade unter den ersten 10 Treffern sortiert. Sogar Retrieval-Strategien, die bewußt stark reduzierte Recall-Werte in Kauf nehmen, führen dennoch oft zu Ballast-Mengen, die eine qualifizierte Relevanzprüfung kaum ermöglichen.



Die obige Grafik illustriert die negative Korrelation zwischen hoher Precision und hohem Recall.[12]

Die folgende Grafik illustriert die Zunahme von Ballast: Bei steigendem (x) steigt (y) ebenfalls proportional an.[13]



Internet-Nutzer müssen ihre Fähigkeiten optimieren, Retrieval-Ergebnisse zu evaluieren. Mit der Zeit haben Bibliothekare und andere *Information-Professionals* Kriterien entwickelt, die nützlich sein können, um die Relevanz von Informationsressourcen einzuschätzen. Nutzer sollten sich verstärkt mit diesen Kriterien vertraut machen.[14] Hier sind einige dieser Kriterien:

- Format - Informations-Ressourcen im Internet werden meist im FTP-, Gopher- oder im HTTP-Format angeboten. Unterschiedliche Verfahren sind erforderlich, um die Verfügbarkeit von Dokumenten in diesen Formaten zu verifizieren. Auch können nicht alle Nutzer auf Anhieb alle Ressourcen-Typen ohne weiteres abrufen, z.B. haben nicht alle kommerziellen Internet-Anbieter standardmäßig Web-Browser im Angebotspaket.
- Reichweite - Die Reichweite einer Informations-Ressource ist ein Maßstab für die Art, wie sie Quellenmaterial behandelt, welche Themen berücksichtigt werden und wie aktuell die Daten sind. Der vorgesehene Adressatenkreis kann die Reichweite ebenfalls stark beeinflussen.

---

[12] Quelle: http://www.csee.umbc.edu/~ian//irF02/lectures/09Evaluation.pdf (akzessioniert am 4. September 2003)
[13] Quelle: http://ir.iit.edu/~nazli/CS495-Slides/01Introduction.pdf (akzessioniert am 4. September 2003)
[14] Smith, Linda C. (1991) "Selection and Evaluation of Reference Sources" aus: Richard E. Bopp & Linda C. Smith (eds.), Reference and Information Services: An Introduction. Englewood, CO: Libraries Unlimited, S. 240

- Beziehung zu anderen Werken - Wie man bei einem offenen, allgemein zugänglichen Netzwerk erwarten würde, das keine inhaltlich arbeitende organisierende oder regulierende Körperschaft besitzt, gibt es im Internet vielfach informatorische Überschneidungen. Wer Lesezeichen (also URL-Sammlungen) von Internet-Ressourcen anlegt, hat in Bälde eine große, ziemlich unübersichtliche Datei. Wer die Beziehungen der Ressourcen untereinander kennt, insbesondere ihre gegenseitigen Verlinkungen, kann seine elektronischen Informationen besser managen. Außerdem haben manche Internet-Ressourcen auch gedruckte Versionen, die möglicherweise sogar mehr Informationen beinhalten können als die elektronischen Versionen. Das Wissen darum unterstützt den Nutzer bei der Bemühung um die Beschaffung von möglichst aktuellen und vollständigen Informationen.
- Autorität - Das Wissen um die Affiliation oder den sonstigen beruflichen Hintergrund eines Ressourcen-Anbieters kann hilfreich sein, um Verläßlichkeit und Pertinenz der Ressource und deren Informationsangebots zu beurteilen. U.a. können persönliche Homepages, Personalverzeichnisse von Hochschulen und virtuelle Nachschlagewerke nützliche Informationen über Informationsanbieter liefern.
- Behandlung - Die Identifizierung des Adressatenkreises einer Internet-Ressource gibt Hinweise auf die Art der intendierten Behandlung, z.B. ob sich die Ressource an ein einschlägiges Fachpublikum oder an die Allgemeinheit richtet. Eine Analyse der Objektivität der Ressource hilft, die Verläßlichkeit des Materials zu beurteilen.
- Anordnung - Das Internet besitzt kein übergeordnetes Organisationsschema oder -struktur. Jedem Ressourcen-Anbieter ist es überlassen, eigene Binnenstrukturen (alphabetisch, klassifikatorisch, dem Organigramm folgend u.ä.m.) oder aber gar keine vorzusehen. Wie gut das Angebot einer Ressource angeordnet ist, wirkt sich auf deren Verwendbarkeit aus.
- Kosten - Obwohl das Internet oft als prinzipiell kostenlose Informations-Ressource bezeichnet wird, gibt es einige versteckte und nicht-so-versteckte Kosten. Zunächst einmal wurde das Internet durch Mittel der öffentlichen Hand in den USA entwickelt; ein Teil der Infrastruktur wird weiterhin öffentlich gefördert. Wer einen eigenen Internet-Zugang haben will, benötigt zumindest einen PC, ein Modem bzw. einen anderen Datenleitungsanschluß sowie einen Internet-Dienstleister. In einigen Gegenden werden Zugänge für Personen, die dort einen Wohnsitz haben, von sog. Freenets kostenlos zur Verfügung gestellt. Dies ist aber eher die Ausnahme. Viele Internet-Nutzer haben über eine akademische oder eine berufliche Affiliation einen Zugang. Zugänge können von kommerziellen Dienstleistern kostenpflichtig erworben werden. Die Kosten hierfür variieren beträchtlich je nach Dienstleister und Umfang des erwünschten Leistungspakets. Und nicht alle Internet-Angebote sind kostenlos; es gibt eine sehr hohe Zahl an entgeltpflichtigen Diensten. Suchzeiten werden je nach ISP auch über die Telefonrechnung abgerechnet. Immaterielle Kosten entstehen in bezug auf die eigene Arbeits- oder Lebenszeit, die man mit Surfen oder gezielten Recherchen verbringt. Für einige Menschen mag dies kein Problem darstellen; für andere kann es schon zu einem Problem werden.

Obwohl diese Kriterien ursprünglich entwickelt wurden, um gedruckte Informationsquellen zu bewerten, und hier nur per Analogieschluß auf Online-Ressourcen übertragen wurden, können sie dennoch auch in diesem Fall als Wegweiser dienen.

Ausgehend vom erstmaligen Retrieval-Ergebnis müssen Nutzer imstande sein, ihre Suchargumente so zu modifizieren, daß die Ergebnisse ihren Erwartungen und Anforderungen besser entsprechen. Techniken wie die Verwendung Boole'scher Operatoren, um gesuchte Sachverhalte in ihre logischen Komponenten zu zerlegen, die Verwendung von spezifischeren oder selteneren Termini, um die Relevanz der gefundenen Dokumente zu steigern, Einsetzung

von Synonymen, um die Suche zu erweitern, die Auswertung von gefundenen Dokumenten nach neuen suchrelevanten Deskriptoren, sowie die Benutzung von anderen Datenbasen oder von anderen Suchmaschinen, um die Suche zu verfeinern und somit den Precision-Wert zu steigern und die Ballast-Menge zu verringern, sind alle in diesem Kontext nützlich.

Es kostet zugegebenermaßen Zeit, um zu lernen, derartige Techniken erfolgreich anzuwenden, aber es kostet ebenfalls Zeit, um immer und immer wieder Hunderte oder Tausende von falschen Treffern zu sichten.

Nichts an der Situation vis-à-vis der Informationsüberflutung wird sich in absehbarer Zeit ändern: Das Internet wird weiterhin exponentiell wachsen, WWW-Seiten werden so konstruiert und indexiert, wie die Informationsanbieter es für richtig halten, Seitenrekrutierung für Suchmaschinen wird mehr als unvollständig sein, indessen große Teile des Webs für die Spinnen, Crawler und 'Bots' der Suchmaschinen unzugänglich bleiben werden. Error 404-Meldungen werden ein alltägliches Ereignis bleiben. Weitere technische Hilfen werden zwar entwickelt, aber sie werden nur palliative Wirkung haben. Informationsüberflutung wird man weiterhin als Problem wahrnehmen, das Streßgefühle verursacht. Precision und Recall werden weiterhin leiden.

Letzten Endes könnte es sein, daß der größte Nutzen des Internets darin besteht, ein gigantisches Branchenverzeichnis zum Auffinden von Experten zu sein. Wenn Benutzer die Personen identifizieren können, die das Wissen oder die Kompetenz haben, die sie suchen, kann man sie dann meist per E-Mail oder sogar offline befragen.

4. Nutzer brauchen mehr Informationskompetenz

Informationskompetenz (engl.: *information literacy*) geht über die bloße EDV-Anwenderkompetenz hinaus. Sie kann einen nennenswerten Beitrag zur Steigerung der Qualität der Retrieval-Ergebnisse von nicht-professionellen Anwendern leisten.

Worin besteht Informationskompetenz und wodurch kann sie erreicht werden?

Nach Feststellungen der American Library Association (ALA) ist Informationskompetenz eine Gruppe von Fähigkeiten, wodurch Individuen imstande sind "to recognize when information is needed and have the ability to locate, evaluate, and use effectively the needed information."[15] Die ALA hat einen Informationskompetenz-Standard[16] entwickelt, nach dem ein informationskompetentes Individuum folgendes kann:

- Feststellen, wieviel Information benötigt wird
- Sich die benötigten Informationen effizient beschaffen
- Information und ihre Quellen kritisch einschätzen
- ausgewählte Informationen in den eigenen Wissensbestand integrieren
- Informationen effizient benutzen, um eine spezifische Aufgabe zu erfüllen
- Die ökonomischen, juristischen und gesellschaftlichen Fragen um die Informationsnutzung verstehen
- Sich die benötigten Informationen ethisch und juristisch einwandfrei beschaffen

[15] American Library Association. "Presidential Committee on Information Literacy. Final Report". (Chicago: American Library Association, 1989.)
[16] ALA Information Literacy Standards. URL: http://www.ala.org/Content/NavigationMenu/ACRL/Standards_and_Guidelines/Information_Literacy_Competency_Standards_for_Higher_Education.htm#stan (akzessioniert am 23. Juni 2003)

Informationskompetenz bildet die Basis für einen lebenslangen Lernprozeß. Sie ist inter- bzw. transdisziplinär, unabhängig von spezifischen Lernumgebungen sowie von einem bestimmten Bildungsstand. Sie setzt einen in den Stand, neue Inhalte zu meistern und seine Recherchen zu erweitern. Dadurch wird man selbstbestimmter und hat mehr Kontrolle über den eigenen Lernfortschritt. Informationskompetenz hat auch mit sicherer IKT-Bedienungstechnik zu tun, hat aber viel weitergehende Implikationen für Individuen, das Bildungssystem und für die Gesellschaft.

5. Fazit

Es gibt viele Ursachen für die Informationsüberflutung, daher gibt es auch kein Patentrezept zur Lösung des Problems, sondern man muß auf eine Vielzahl von Remeduren setzen. Obwohl es mittlerweile eine Fülle von Lehr- und Lern-Ressourcen für die Vermittlung von Informationskompetenz gibt, steht zu vermuten, daß die nachhaltige Optimierung von Internet-Retrievalergebnissen von nicht-professionellen Anwendern im wesentlichen auf dem Selbst-Lernprinzip basieren muß. Auf jeden Fall sollte man keine allzu starken Hoffnungen in technikbasierte Lösungsansätze setzen.