Suchmaschinenlandschaften

Klaus Patzwaldt

Wer spontan nach Namen von Suchmaschinen fragt, bekommt außer Google und Yahoo! meist nur noch wenige andere Anbieter genannt. Am mangelnden Angebot kann dies nicht liegen, denn Anfang 2005 existierten laut klug-suchen.de allein 1.600 meist deutschsprachige durchsuchbare Datenbestände. Das Portal Sucharchiv.com verzeichnete rund 4.700 Suchmaschinen, Archive und Portale zum Auffinden von Informationen im Internet. Diese Vielfalt mag signalisieren: Suchmaschinen gehören zum Internet wie das Salz zur Suppe; doch ist ihr Verhältnis untereinander nicht ganz unproblematisch: Suchmaschinen sind zwar fast so alt wie das World Wide Web selbst, doch können sie bisher mit seiner rasanten Entwicklung nicht mithalten. Lange Zeit wurden multimediale Inhalte von wichtigen Suchmaschinen nicht beachtet. Es wurden ausschließlich Texte indexiert. Und selbst wenn heute multimediale Inhalte erfasst werden, gelangen diese lediglich in Textform in den Index der Suchmaschinen. Verglichen mit der menschlichen Entwicklung können wir also von einer embryonalen Phase sprechen, in der sich die Suchmaschinen befinden.

Kurze Geschichte der Suchmaschinen

Alles begann mit *Archie* (Abkürzung von »Archive«), einer Suchmaschine die von Alan Emtage an der McGill-Universität in Montreal entwickelt wurde und zwischen 1990 und 2000 aktiv war. Das System durchsuchte damals noch nicht das WWW, sondern lediglich FTP-Server, um Daten und Dokumente zu katalogisieren. Die erste Suchmaschine für das Web hieß *The Wanderer* und wurde vom MIT-Student Matthew Gray entwickelt. The Wanderer zählte zunächst nur vorhandene Webserver, erst später wurden auch die Webadressen erfasst: Im Juni 1993 zählte The Wanderer überschaubare 130 Webserver (=Websites), im Dezember 1994 waren bereits 10.022 Einträge im Index.

Im Januar 1994 begannen dann die beiden Studenten David Filo und Jerry Yang Suchmaschinen-Geschichte zu schreiben. Die beiden erstell-

ten ein Verzeichnis Ihrer beliebtesten Webadressen, gründeten im März 1995 das Unternehmen *Yahoo!* und boten von Beginn an ein strukturiertes, redaktionelles Verzeichnis an, welches z.B. The Wanderer überlegen war, weil es zu jeder Webadresse eine Beschreibung gab. Konkurrent *Lycos* startete im Juli 1994 mit einem Katalog von 54.000 Dokumenten und zählte im April 1995 bereits 2,95 Millionen Einträge im Index. Lycos begann sowohl die Webadresse als auch die ersten Zeilen zu speichern. Für die Häufigkeit des Suchbegriffes wurde damals eine Trefferquote ermittelt, um die Suchergebnisse zu verbessern.

Am 15. Dezember 1995 stellte der Computerspezialist Digital Equipment Corporation (DEC) die Suchmaschine *AltaVista* vor. Diese kannte im Mai 1996 mehr als 30 Millionen Webseiten und drei Millionen Usenet-Postings [@1]. AltaVista indexierte jedes Wort von jeder Webseite, startete also als Volltextsuchmaschine. AltaVista bot als erster Anbieter eine erweiterte Suche, Linkprüfung und das Hinzufügen einer Seite in den Index innerhalb von 24 Stunden. »Super Spider« und »Super Indexer« wurden auf damals revolutionären 64-Bit-Alpha-Servern betrieben. Der aktuelle, umfangreiche Index machte AltaVista für einige Jahre zur führenden Suchmaschine.

Als *Google* offiziell am 7. September 1998 startete, blieb der breiten Öffentlichkeit das Potenzial der Treffergenauigkeit dieser Suchmaschine verborgen. Während sich AltaVista das eigene Grab schaufelte, indem immer mehr Portalfunktionen den Suchschlitz immer unauffindbarer machten, fanden und finden Google-Nutzer eine schlichte, klar gestaltete Suchseite vor. Die PageRank-Technologie verhalf Google zudem zu einem Erfolg, den bisher keine andere Web-Suchmaschine für sich verbuchen konnte. Als Basis diente die nach Ansicht der Erfinder Larry Page und Sergey Brin demokratische Struktur des Internet: Jeder Verweis auf eine Seite wird als eine Wahlstimme betrachtet.

Geschichte deutschsprachiger Suchmaschinen

Der deutsche Suchmaschinenmarkt brachte zwar eine ganze Reihe von Suchhilfen hervor, doch nur wenige waren wirklich interessant. Viele dieser Suchmaschinen waren Eigenentwicklungen von Einzelpersonen oder kleinen Unternehmen, die mit dem Fortschritt der großen Suchmaschinen nicht mithalten konnten. Eine der wenigen beachtenswerten Entwicklungen war *Fireball*: Sie startete im Juni 1997 als Ausgründung der TU Berlin, wurde später von *Lycos* übernommen und liefert seit dem Frühjahr 2004 keine eigenen Ergebnisse mehr. Ein weiteres interessantes Projekt: *Infoseek Deutschland* ging im April 1999 als Joint Venture an den Start und schaffte es nach kurzer Zeit schwarze Zahlen zu schreiben. Entsprechend einer Gesellschafterentscheidung von T-Online wurde In-

foseek Ende Oktober 2001 aber wieder eingestellt. Die ehemaligen Infoseek-Geschäftsführer starteten im Frühjahr 2004 mit der Suchmaschine *Seekport* einen zweiten Anlauf. Seekports Datenbank wird redaktionell von einem Index-Team betreut und will gegen Spam, Dialer und andere Plagen mit menschlichem Verstand vorgehen.

Trotz aller Anläufe, Ideen und Millionen Euro von Venture Capital: Marktbeherrschend waren im Jahr 2004 nur zwei Suchmaschinenbetreiber: Google und Yahoo! – verfolgt vom Microsoft-Ableger MSN, der Anfang 2005 mit neuem Angebot und eigener Technologie startete.

Finanzierung von Suchmaschinen

Alle derzeit erfolgreichen Suchmaschinen arbeiten auf kommerzieller Basis. Die Nutzung der Suchmaschinen wird kostenfrei angeboten, die Finanzierung erfolgt daher auf anderen Wegen. Google finanziert sich entsprechend der zum Börsengang gemachten Angaben (im Frühjahr 2004) zu über 95 Prozent aus Werbung. Sie wird auf den Ergebnisseiten der Suchmaschine und im Partnernetzwerk eingeblendet. Traditionelle Finanzierungsmodelle beruhen auf der Lizenzierung der Suchmaschinentechnologie zum Einsatz in Firmennetzwerken sowie auf der Lizenzierung zum Gebrauch des Suchmaschinenindexes auf anderen Webseiten. Damit wird aber nur ein verhältnismäßig geringer Beitrag zum Gesamteinkommen erzielt. Yahoo! definiert als eigene Aufgabe, wichtigster Internetservice für Privatpersonen und Unternehmen weltweit zu werden. Für diesen Zweck sind alle Aufgaben recht: Entsprechend ist die Internet-Suche nur ein Teil der globalen Yahoo!-Strategie, gehört jedoch zu den wichtigsten Bereichen im Unternehmen. Für die Werbevermarktung ist das Tochterunternehmen Overture zuständig. Yahoo! setzt zunehmend auf den Vertrieb kommerzieller Services wie z.B. DSL-Internet-Zugang, und sichert sich damit kontinuierliche Einnahmen, die von der Werbung unabhängig sind.

Freie Suchmaschinen

Die Idee freier Suchmaschinen orientiert sich an Beispielen, die zeigen, wie erfolgreich freie Software sein kann. Dazu gehört z.B. das freie Betriebssystem Linux, der Apache-Webserver oder die Enzyklopädie Wikipedia. Das Open Directory Project (ODP) zeigt, wie sich eine Orientierungshilfe für das Internet etablieren kann: Mehr als 66.000 freiwillige Helfer aus der ganzen Welt arbeiten in einem redaktionellen System zur

Beurteilung von Websites. Sehr viele kleinere, kommerzielle Webverzeichnisse (wie Dino-Online) haben ihren Dienst wieder eingestellt, weil sich der redaktionelle Aufwand nicht finanzieren ließ und bieten nun die Inhalte des ODP an. Ein ähnlicher Erfolg blieb freien Suchmaschinen bisher versagt.

Der Erfolg freier Suchmaschinen steht und fällt mit einer starken Entwickler- und Anwendergemeinde, denn Suchmaschinen sind einer sehr intensiven technischen Entwicklung unterworfen und können nur dauerhaft bestehen, wenn sie sich ständig weiterentwickeln. In Deutschland wurde im Juli 2004 der »Verein zur Förderung der Suchmaschinentechnologie und des freien Wissenszugangs« (SuMa-eV) [@2] gegründet. Der Verein hat das Ziel, den freien Zugang zum Wissen unabhängig von kommerziellen Interessen zu gewährleisten. Dafür soll der Aufbau einer dezentralen, kooperativen und nicht-monopolistischen Suchmaschinenstruktur zunächst in Deutschland geplant und durchgeführt werden. Eine Rolle, in der die Gründer des SuMa-eV viel lieber den Staat sehen, der nach dem Modell der öffentlich-rechtlichen Anstalten die Unabhängigkeit der Suchmaschinen gewährleisten könnte. Rein technisch dürfte eine dezentrale Struktur die beste Voraussetzung für eine globale Suchmaschine sein, weil die Kosten für Hardware oder Bandbreite von den Nutzern getragen werden. Nach dem P2P-Prinzip (Peer-to-Peer) unter Beteiligung vieler Freiwilliger arbeitet z.B. das Seti@home-Projekt [@3], das sich der Suche nach Signalen außerirdischer Wesen verschrieben hat. Ähnlich könnten Beteiligte mit ihrem Computer ständig einen kleinen Teil des Internet auf neue und geänderte Seiten überprüfen.

Entwickler freier Suchmaschinen argumentieren, dass kommerzielle Dienste von Interessengruppen kontrolliert würden. Die einzige Möglichkeit unvoreingenommene Ergebnisse zu erzielen, wäre deshalb die Verwendung von öffentlich zugänglicher Software. Diese Argumentation vergisst aber, dass die Offenheit einer Technologie die Möglichkeiten der Manipulation erleichtert. Diese Manipulation der Suchergebnisse war und ist eine sehr viel stärkere Bedrohung für Suchmaschinen als die angebliche und tatsächliche Zensur durch fremde Mächte. Da wir hier vor allem Suchmaschinen betrachten, die mit Google, Yahoo! und MSN konkurrieren sollen, müssen entsprechende technische Grundlagen vorhanden sein. Die zu verwaltenden Datenmengen bewegen sich bereits jetzt im Petabyte-Bereich. Bisher gibt es keine freie, nichtkommerzielle Suchmaschine, die praktisch auch nur ansatzweise eine ernsthafte Konkurrenz zu den Globalplayern darstellt. Probleme, die es zu bewältigen gibt, sind zahlreich: ausreichende technische Ressourcen, erfolgreicher Umgang mit Suchmaschinen-Spam, betrügerischer Manipulation und Missbrauch der Anwendung.

Das ambitionierteste Projekt für freie Suchmaschinen ist Nutch, geleitet von Doug Cutting, das jedoch keine eigene Suchmaschine betreibt: Die von Nutch [@4] bereitgestellte, im Quelltext offene Software, kann in beliebigen kommerziellen und nichtkommerziellen Suchmaschinen verwendet werden. Yacy [@5] arbeitet nach dem oben erwähnten P2P-Prinzip. Vorausgesetzt, genügend Nutzer wären bereit, sich die notwendige Software auf ihrem PC zu installieren, könnte eine globale Suchmaschine mit einem dezentralen Index geschaffen werden. Die Gefahr, Manipulationen nicht zu beherrschen, erscheint hier jedoch noch größer, weil kein zentrales Verwaltungsinstrument vorgesehen ist, welches solche Versuche unterbinden könnte. Murray Walker beschreibt auf minty.org [@6] weitere interessante Gedanken zum Betrieb offener, verteilter Suchmaschinen für das Internet.

Stärken & Schwächen von Suchmaschinen

Suchmaschinen sind manipulierbar: Durch Betreiber und Nutzer. Die zunehmende kommerzielle Natur des Internet weckt Begehrlichkeiten für die Platzierung in Suchmaschinen. Nutzer sind bequem und nur selten bereit, ihre Nachforschungen über die erste Ergebnisseite auszuweiten. Das bedeutet, wer nicht auf der ersten (eventuell noch zweiten oder dritten) Seite platziert ist, existiert praktisch nicht. Je höher der finanzielle Anreiz ist, durch geschickte Platzierung in Suchmaschinen neue Kunden zu gewinnen und somit höhere Umsätze zu erreichen, desto stärker ist der Wettbewerb. In Branchen wie Reise, Versicherung, Immobilien oder Automotive kämpfen beauftragte, professionelle Suchmaschinenoptimierer darum, ihre Kunden auf die vordersten Plätze zu bringen. Doch längst sind es nicht nur diese Branchen und nicht immer steht der direkte Verkauf im Vordergrund. Mitunter geht es einfach um das Image: Man möchte vor den Mitbewerbern platziert sein. Zudem haben Untersuchungen ergeben, dass Nutzer Firmen als renommiert und vertrauenswürdig ansehen, wenn sie auf vorderen Plätzen zu finden sind. Speziell kleinere Unternehmen können davon profitieren und ihren Firmennamen durch gute Platzierungen stärken. Suchmaschinenbetreiber halten ihre Ranking-Algorithmen zwar geheim um Manipulationen zu erschweren, verhindern können sie die Manipulationen jedoch nicht.

Die globale Ausrichtung der marktführenden Suchmaschinen sowie die Präsenz von Niederlassungen in zahlreichen Ländern weltweit veranlassen Suchmaschinenbetreiber, sich mit den nationalen ethischen, moralischen und gesetzgebenden Bedingungen auseinanderzusetzen. Während in Deutschland die Toleranz gegenüber nationalsozialistischen In-

halten wesentlich geringer ist als in den USA – sie gelten dort als freie Meinungsäußerung – gibt es in den USA weniger Toleranz gegenüber erotischen Inhalten. Entsprechend dieser und sehr vieler anderer nationaler Eigenheiten gilt es die Inhalte so zu präsentieren, dass keine nationalen Belange verletzt werden. Die Grenzen zur Zensur sind dabei fließend: Nationale Vorschriften verlangen z.B. in Frankreich und Deutschland den Zugriff auf nationalsozialistische Inhalte einzuschränken. Kritiker bemängeln diese Unterdrückung von Suchergebnissen ebenso wie die weit reichenden Einschränkungen der Google-Ergebnisse in der chinesischen Version, einschließlich der chinesischen Ausgabe der Google News. Doch es gibt andere, notwendige Restriktionen der Suchmaschinenbetreiber: Website-Betreiber aus dem Rotlichtbereich sind dafür bekannt, Webseiten zu manipulieren, um für möglichst viele Begriffe auf den vorderen Plätzen zu landen. Die Verbannung von jugendgefährdenden Treffern aus den Ergebnislisten ganz allgemeiner Anfragen dient einem gesellschaftlichen Interesse, denn die wenigsten Eltern sind in der Lage Kindern bereits im Vorschulalter die notwendige Medienkompetenz für den problemlosen Umgang mit dem Medium Internet zu vermitteln. Zudem werden thematisch unpassende Treffer als Belästigung empfunden und schränken die Nutzbarkeit der Suchmaschine ein. Suchmaschinenbetreiber können also ebenfalls die Ergebnisse manipulieren und tun dies auch ständig. In der Regel, um die Ergebnisse im Sinne der Nutzer ständig zu verbessern. Mitunter auch, um sich wirtschaftlichen oder politischen Interessen zu unterwerfen. Das wirft die Frage auf, wie unabhängig Suchmaschinen sein können. Ob staatlich geführte Suchmaschinen unabhängiger sein können und quelloffene Suchmaschinen gegen Manipulationsversuche und den Einfluss der Öffentlichkeit immun sind, muss sich zeigen. Ist es notwendig oder erwünscht, Suchmaschinen zu betreiben, die völlig ohne jegliche redaktionelle Kontrolle arbeiten?

Die Größe des Datenmeeres

Das Internet ist eine völlig inkonsistente Ansammlung von Webseiten, deren Anzahl in jeder Minute andere Werte annimmt. Neue Seiten werden bereitgestellt, ältere Dateien entfernt. Es sind keine öffentlichen Daten zur Anzahl vorhandener Webseiten bekannt. Lediglich die ständig steigende Anzahl der aktiven Server wird von Netcraft [@7] ermittelt. Generell steigt die Anzahl der erfassten Seiten kontinuierlich. Doch nicht alle Veröffentlichungen sollen in den Index der Suchmaschinen. Kommerzielle Anbieter sorgen z.B. dafür, dass ihre Inhalte (z.B. Wirtschafts- und Rechtswissen), nur soweit gefunden werden, wie es zum Anlocken

neuer Kunden notwendig ist. Weitere Lücken entstehen durch technische Probleme, wie bei Webseiten, die dynamisch aus Datenbanken generiert werden. Zahlreiche Angebote, die mit einem Content-Management-System (CMS) erstellt wurden, Webshops sowie generell Angebote, die Interaktion benötigen, können ohne zusätzliche technische Maßnahmen nur beschränkt oder gar nicht erfasst werden. Nach einer Studie von Brightplanet [@8] ist das für Suchmaschinen »unsichtbare Web« (Deep Web) ca. 400- bis 550-mal größer, als das »sichtbare Web« (Surface Web). Zahlreiche Angebote hingegen werden ausschließlich dafür produziert, in Suchmaschinen an vorderen Stellen gefunden zu werden. Sie leiten auf andere, kommerzielle Angebote weiter ohne selbst einen Nutzeffekt aufzuweisen.

Neue Webseiten kommen mit einer Verzögerung von mehreren Tagen bis Monaten in den Index der Suchmaschinen. Die Nutzer finden also kein aktuelles Abbild des Internet vor, sondern lediglich ein Abbild aus der Vergangenheit. Dieses Problem wird teilweise kompensiert, indem oft aktualisierte Webseiten häufiger besucht werden.

Die Nutzer sollten bedenken, dass Suchmaschinen nur teilweise gleiche Bereiche des Internet abdecken und mit unterschiedlichen Rankingkriterien arbeiten. Was bei einer Suchmaschine nicht auf vorderen Plätzen sichtbar ist, finden Sie möglicherweise in einer anderen Suchmaschine. Allgemeine Suchmaschinen wie Google und Yahoo! sind längst nicht die einzigen Anlaufpunkte für die Recherche. Viele fachspezifische Informationen sind ausführlicher über spezielle Datenbanken zu erreichen, z.B. »GEIN« [@9] für Umweltthemen.

Aussichten

Maschinenbasiertes Lernen und künstliche Intelligenz sind wesentliche Einflussfaktoren, die den zukünftigen Fortschritt der Suchmaschinen bestimmen werden. In der weiteren Entwicklung werden Suchmaschinen entstehen, die den Eindruck erwecken über menschliche Eigenschaften zu verfügen und »charakterstark« zu sein. Das wichtigste Merkmal einer Suchmaschine wird ihre Kommunikationsfähigkeit sein. Wurde eine Frage nicht oder nicht ausreichend verstanden, wird die Suchmaschine in einem »freundschaftlichen« Gespräch nachfragen. Schrittweise wird die Fragestellung eingegrenzt, um eine möglichst präzise Antwort zu ermöglichen. Wenn Nutzer es zulassen, kann die Suchmaschine aus dem bisherigen Suchverhalten Rückschlüsse ziehen. Wir können zudem Interessengebiete angeben, die gleich dafür sorgen, dass Suchergebnisse für »Jaguar« Informationen aus dem Tierreich enthalten sollen, nicht aus

dem Bereich »Autos«. Der Umgang mit persönlichen Daten (Personalisierung) wird auch zukünftig auf große Skepsis stoßen, deshalb kann eine Suchmaschine zukünftig auch ohne vorheriges Datensammeln notwendige Informationen zum Eingrenzen der Suchanfrage in einem intelligenten Dialog abfragen.

»Endlose« Ergebnislisten wird es prinzipiell weiterhin geben. Diese werden jedoch nur von Nutzern abgefragt, die sich wirklich durch zahlreiche Informationen durcharbeiten möchten oder müssen. Ein gewöhnliches Ergebnis wird jedoch sehr präzise einen oder eventuell ganz wenige weitere Treffer liefern. Die Ermittlung der Relevanz eines Treffers wird sich zukünftig viel stärker an den Kontext der Informationen anlehnen, als es die bisherige Worterkennung ermöglicht. Die größte Herausforderung ist die ständig wachsende Masse unstrukturierter Daten im Internet und deren kontextuelle Auswertung. Dabei geht es unter anderem um die Erkennung verschiedener menschlicher Ausdrucksweisen. Neben sachlicher Argumentation können das z.B. Ironie oder Humor sein.

Nahezu parallel werden Software-Agenten entwickelt, die Aufgaben ihrer Nutzer weitgehend selbstständig erledigen – unter anderem die Suche von Informationen mit Hilfe von Suchmaschinen und anderer Quellen im Internet. Der Agent wird vor allem dabei behilflich sein, für Suchanfragen die passende Suchmaschine (lokale bzw. Spezialsuchdienste) auszuwählen und qualifizierte Anfragen an die Suchmaschinen zu senden. Die bisherigen Suchanfragen des Nutzers werden dafür ausgewertet. Die Ergebnisse können vom Agenten durch die Kenntnis der persönlichen Vorlieben des Nutzers zusätzlich vorsortiert werden. Weil Agenten als persönliche Software auf dem Computer der Anwender installiert werden, kann die Privatsphäre weitgehend geschützt bleiben, denn die gesammelten Informationen verlassen nicht den Rechner.

Digitale Verweise

[@1] web.archive.org/web/19960511013133/http://www.altavista.digital.com
[@2] www.suma-ev.de
[@3] http://setiathome.ssl.berkeley.edu/
[@4] www.nutch.org
[@5] www.yacy.net
[@6] http://search.minty.org
[@7] www.netcraft.com
[@8] www.brightplanet.com
[@9] www.gein.de