



Internet-Suchmaschinen: wie kommen sie zu ihren Daten?

Ein Vortrag im Rahmen der AKI Hamburg am 23. Februar 2006

Inhalt:

- Wie sehen die Informationen des Internet aus Sicht einer Suchmaschine aus?
- Woher können Suchmaschinen qualifizierende Aussagen zu den Inhalten gewinnen?
- Was für Informationen werden erfasst und welche nicht?
- Was erwarten die Benutzer von Suchmaschinen und was wird real erfüllt?
- In wie weit sind die Ergebnislisten und die Inhalte von Suchmaschinen manipulierbar?
- Was sind die Geschäftsmodelle von Suchmaschinen?
- Wo wird die Technologie aktuell weiterentwickelt?



1 Einleitung

- Die Recherche von Information im Internet gewinnt immer mehr an Bedeutung.
- Das Suchverhalten der Nutzer hat sich verändert.
- Suchmaschinen kommt eine zentrale Makler-Funktion zu.

Für den verantwortungsvollen Umgang mit diesen Informationen ist ein grundlegendes Verständnis für die zu Grunde liegende Technologie und der daraus resultierenden Probleme notwendig:

- technische Probleme: Datenvolumen, Doubletten, unklare Datenquelle, Werbemüll,
- politische Probleme: Monopolisierung, Bewertung und Filterung von Inhalten, ...

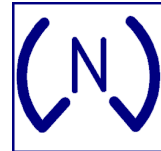
2 Suchverhalten im Internet

"Klassischer" Ansatz:

- Lexikon, Bibliothek, Dienstleister
- Hochwertiger und relativ vollständiger Datenbestand
- Strukturierte Fragen
- Qualitativ hochwertige Antworten
- Antwortzeit Minuten bis Tage

Internetsuche:

- "googlen" / wikipedia
- Vom Nutzer autodidaktisch gelernt
- Umfangreicher Datenbestand von sehr heterogener Qualität
- Ausgehend von wenigen Worten wird der Ergebnisraum iterativ unterteilt.
- Beachtung nur der ersten Ergebnisse.
- Reaktionszeit Sekunden.



3 Eine perfekte Suchmaschine?

- *Vollständige Erfassung der verfügbaren Informationen*

Eine Suchmaschine im Internet soll "das (ganze) Internet" durchsuchen. Je nach der Erfahrung des Nutzers mit dem Medium vielleicht auch nur "den wichtigsten Teil des Internet".

- *Hochwertige Ergebnisse*

Die Qualität der Ergebnisse soll einen Bezug zur Frage besitzen und objektive Antworten liefern.

- *Einfachheit bei der Benutzung*

Die Suchmaschine soll bereits mit nur wenigen Stichworten das erwartete Ergebnis liefern. Die Fragen werden selten ausformuliert sondern durch eine Liste von Stichworten umschrieben.

- *Hohe Aktualität der Informationen*

Die der Suche zu Grunde liegenden Informationen sollen dem aktuellen Abbild der Quelle entsprechen. Änderungen an der Quellseite gehen ohne Verzögerung in die Suche mit ein.

- *Objektive Behandlung der Inhalte*

Eine Suche wird in der Regel mehr als ein Ergebnis liefern. Diese sollen so gewichtet werden, dass das "beste Ergebnis" als erstes geliefert wird.

4 Informationsquellen im Internet

Woher kommt das "Wissen" des Internets?

Strukturierte Quellen	Dynamische Quellen
<ul style="list-style-type: none">• Redaktionelle Inhalte, Medien• Hersteller• Forschungseinrichtungen• Öffentliche Einrichtungen	<ul style="list-style-type: none">• Newsgroups• Mailinglisten• Private Homepages• Vereine• Foren• Blogs
Merkmale: <ul style="list-style-type: none">- bekannte Quellen- für Suchmaschinen optimiert- von Suchmaschinen (oberflächlich) erfasst	Merkmale: <ul style="list-style-type: none">- ungeordnete Bereitstellung- "subjektiv" aber hochwertig- schwierig zu erfassen



5 Grenzen von Suchmaschinen

- *Datenmenge*

Der Umfang der Ausgangsdaten ist kaum abschätzbar. Kontinuierlich werden neue Informationen erzeugt und alte verschwinden. Die Qualität der Ausgangsdaten wird durch Werbemüll und Webspam gezielt verringert.

- *Aktualität der Datenbasis*

Eine Suchmaschine muss die Ausgangsdaten präventiv erfassen und vorverdichten. Zwischen der Erfassung und der Anfrage verstreicht immer eine gewisse Zeitspanne, so dass die Ergebnisse der Suchmaschinen in der Regel veraltet sein müssen.

- *Verfügbare IT-Technik*

Die zu verarbeitenden Datenmengen und Anforderungen an Antwortzeiten und Verfügbarkeit stellen die aktuelle Hard- und Software vor grosse Herausforderungen.

- *Doubletten/Herkunft*

Identische oder stark ähnelnde Inhalte sind oft an verschiedenen Stellen im Internet verfügbar. Die Originalquelle ist nur noch selten erkennbar. Eine Bewertung der Information an Hand des Autors ist nicht mehr möglich.

- *Mehrdeutigkeit der Suchanfragen*

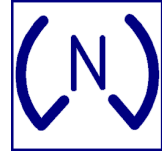
Eine Recherche "per Stichwort" ist ein iterativer Prozess, bei dem sich die Benutzer schrittweise an die erwartete Antwort herantasten. Der Informationsgehalt der Frage ist oftmals für eine korrekte Antwort zu gering.

- *Bewertung der Relevanz*

Die Relevanz einer Antwort auf eine Frage hängt häufig auch von der Umwelt des Fragenden ab. Die Definition einer objektiven Relevanz ist auf Grund der fehlenden Rahmendaten nur schwer möglich.

- *Juristische Forderungen*

Suchmaschinen sind Teil einer Gesellschaft und müssen sich den Regeln/Gesetzen der Gesellschaft ihres Heimatlandes unterwerfen. Im Rahmen des Internets gelten diese Regeln aber oft nur für einen Teil der Benutzer. Der andere Teil kennt die ursprünglichen Regeln und ihren Einfluss auf die Suchmaschine und ihre Ergebnisse nicht.



6 Funktionseinheiten einer Suchmaschine

Komponente	Beschreibung
Crawler / Fetcher	"Automatisierter Webbrowser", sammelt Dokumente
Parser	Extrahiert die textuelle Information aus den Dokumenten
Indexer	Erzeugt einen Suchindex für die schnelle Abfrage
Suchfrontend	Interaktion mit dem Benutzer
Suchindex	Zuordnung Worte zu Dokumenten
Dokumenten-Cache	Kopie der eingesammelten Dokumente
Web-Datenbank	Verzeichnis der gesammelten und gefundenen Seiten

7 Wie werden Websites gefunden?

Suchmaschinen müssen selber neue Websites finden und in ihren Index aufnehmen.

Initiiert vom Websitebetreiber:

- Eintragung von URLs in Verzeichnissen ([DMOZ](#), ...)
- Direkte Anmeldung von URLs bei den Suchmaschinen

Automatisch von der Suchmaschine:

- Suche nach ausgehenden Links
- Domainlisten, IP-Nummernlisten
- Manuelle Einbindung wichtiger Websites



8 Grenzen des Crawlers: Deep-Web

Verschiedene Teile des WWW lassen sich nur schwer mit dem automatisierten Webclient (Crawler) erfassen. Dieser Teil wird oft als "deep-", "dark-" oder "invisible web" bezeichnet - kurz was Suchmaschinen nicht sehen (Der Begriff wurde schon 1994 von Dr. Jill Ellsworth geprägt. (siehe <http://www.brightplanet.com/deepcontent/>)).

- Suchmaschinen folgen Verweisen von Seite zu Seite
- oftmals nur die "obersten" Ebenen indiziert (Surface Web)
- Erfasst deutschsprachige Seiten der grossen Suchmaschinen ~100 Mio. Seiten (geschätzt!)
- Umfang des deutschsprachigen Netzes: über 1 Mrd. Seiten (geschätzt)
- nicht zugängliche Inhalte oftmals sehr hochwertig

9 Problematische Seiten

Problematische Bereiche für Crawler sind:

- Fehlerhafte Webseiten (Link-Loops, ungültige Links, Crawler-Traps, ...)
- Gesperrte Webseiten
- zu große Anzahl an Dokumenten auf einer Website (ohne Priorisierung)
- zu langsames Antwortverhalten des Webservers
- unbekannte Websites
- Datenbanken (Eingabefelder)

In der Regel ist es ein Problem der **Websitebetreiber**, ihre Inhalte für Suchmaschinen zugänglich zu machen.



10 Daten aus Sicht des Parsers

Nachdem die Daten über den Crawler eingesammelt worden, müssen sie analysiert werden. Die Suche setzt in der Regel auf Schlagworten auf. Die Aufgabe des Parsers ist die Überführung der Rohdaten in einen indizierbaren Text.

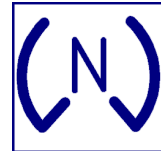
Rohdaten -> Dekodierung -> Text-Extraktion, Feld- Erkennung -> Indizierbarer Rohtext
Link-Erkennung Link-Liste für den Crawler
-> ->

11 Zusätzliche Informationsquellen

Die automatisch erzeugten Daten reichen in der Regel alleine nicht aus. Die Informationen stehen in einem gewissen Kontext, der auch bei der Indizierung beachtet werden muss.

Weitere Informationsquellen sind:

- Analyse der Linkstruktur (Stichwort: Page-Rank)
- Zusätzliche Felder durch die Autoren (Meta-Tags, Dublin-Core-Tags)
- Zuordnung der Website zu Kategorien in Internet-Verzeichnisse (Dmoz, Yahoo)
- Manuelle Kennzeichnung
- Kennzeichnung durch Communities (z.B. Yahoo-Groups)
- Nutzung von Thesauri, Lexika
- Grammatikalische Analyse der Sprache



12 Geschäftsmodelle für Suchmaschinen

Der Betrieb einer Suchmaschine kostet Geld. Wie können die Ausgaben finanziert werden?

- Banner-Werbung
- Einbindung von contextbezogener Werbung
- Platzierung von "Sponsored Links" in den Ergebnislisten
- Beratung und Projekte
- Lizenzen, Produkte
- Individualisierte und zielgruppenbezogene Werbung / Ergebnisseiten

Daneben gibt es verschiedene Suchmaschinen, die von Forschungseinrichtungen, Vereinen oder im Rahmen des Unternehmensimages betrieben werden.

13 Marktanteile

	3.7.2003	7.6.2004	11.10.2005	Umsatz 2004	Umsatz 2005
Google	67.1%	75.1%	83.1%	3,2 Mrd USD	6,2 Mrd USD
Yahoo	7.1%	5.8%	4.2%	3,5 Mrd USD (inkl Overture)	5,3 Mrd USD
MSN Web-Suche	7.7%	5.1%	4.1%	(2,2 Mrd USD)	(2,2 Mrd USD)
AOL Suche	1.5%	2.7%	2.5%	-	-
T-Online	3.4%	2.4%	1.6%	-	-
Altavista	1.4%	0.9%	0.4%	(Yahoo)	(Yahoo)
WEB.DE	1.3%	1.5%	0.5%	43 Mio. Euro	(United Internet: 420 Mio. Euro)
MetaGer	1.5%	1.2%	0.4%	-	-
Lycos	2.1%	1.0%	0.4%	103 Mio. Euro (Europa)	-

Quelle: <http://www.webhits.de/deutsch/index.shtml?/deutsch/webstats.html>, Pflichtveröffentlichungen

Zum Vergleich:

- Axel Springer AG: Umsatz 2005: 2,4 Mrd. Euro
- Gruner + Jahr AG & Co. KG: Umsatz 2004: 2,4 Mrd. Euro



14 Von Google lernen

<u>Frontend</u>	<u>Finanzierung</u>	<u>Technologie</u>
<ul style="list-style-type: none">• Schnelle Antworten• Einfaches Benutzerfrontend• Großer Datenbestand	<ul style="list-style-type: none">• Adwords• Vermarktung in eigener Hand• Auktionsmodell bei Anzeigen	<ul style="list-style-type: none">• Page Rank• "Map-Reduce"-Algorithmus• Personalisierung

Aus Schwächen lernen:

- Suche nach "failure"
- String Suche "Martin Gutschke" (mit und ohne ", ohne Vornamen)

=> Problem: die Wahrung der Datenkonsistenz.

15 Aktuelle Entwicklungen

- "Communitybasiertes Tagging"
- Lokalisierung der Suche (Werbung)
- Semantic Web
- Automatische Übersetzung
- Multimediasuche
- Erfassung von Bibliotheken
- Produktsuche / Vergleiche
- Desktopsuche

16 Bessere Suchmaschinen?

Typische Suchanfragen haben nur sehr wenig Suchbegriffe ("Einwortsuche")

Problem: Doppelbedeutungen von Suchwörtern

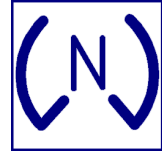
Ziel: automatische Hinführung zur "subjektiv richtigen" Begriffswelt

Ansatz:

- Suchmaschine "kennt" den Benutzer
- Bildung eines Suchprofils (explizite Angabe oder Beobachtung des Verhaltens)
- Programmiertechnisch/logisch einfach, aber HOHER Programmieraufwand

=> HOHER kommerzieller Wert!

=> Alle großen Suchmaschinen arbeiten verstärkt daran!



17 Suchmaschinenoptimierung (SEO)

- Das Internet ist oft ein weiterer wichtiger Vertriebskanal.
- Die Positionierung auf den Ergebnisseiten der grossen Suchmaschinen hat direkten Einfluss auf den Geschäftserfolg.

=> Die Präsentation und Position in den Suchmaschinen ist für die Websitebetreiber wichtig.

- Suchmaschinen finden nicht immer von sich aus die "richtigen" Inhalte
- Eine direkte Einflussnahme auf die Suchmaschinen und die Positionierung in den Ergebnislisten ist nur bedingt möglich.

=> neues Beraterfeld: "Suchmaschinenoptimierung"

18 Werbung und Webspam

Werbung auf Websites

- Früher wurde "Tausender-Kontakt-Preise" bezahlt, d.h. für jede Darstellung einer Werbung erhielt der Websitebetreiber Geld. Dies ist nur noch selten der Fall.
- Aktuell: Die Vermarkter (Amazon, Ebay, Google Adsense, Overture u.v.a.m) zahlen dem Betreiber der Website je Klick auf eine Anzeige einige Eurocent

=> Auf diesem Weg finanzieren sich viele Internetangebote

Webspam

- Webspammer bauen automatisiert Internetseiten (Webspam-Seiten)
- Die Webspam-Seiten sind mit Anzeigen versehen
- Üblicherweise werden alte aber etablierte Domains für solche Angebote aufgekauft
- Suchmaschinen erfassen diese Seiten und geben sie als Ergebnisse mit aus
- Ein kleiner Prozentsatz der Suchenden wird so zu den Spam-Sites geleitet
- Ein noch kleinerer Prozentsatz der Besucher klickt dann auf eine Anzeige

=> in der Menge verdienen grosse Spammer darüber mehrere Tausend Euro je Monat



19 Ergebnis-Manipulation

Die aktuellen (dubiosen) Tricks zur Manipulation der grossen Suchmaschinen sind:

- Cloaking: Unterschiedliche Inhalte für Besucher und Suchmaschinen
- Link-Farm: ermöglichen eines Page-Rank
- Guestbook-Spamming: ermöglichen eines Page-Rank durch automatische Eintragung
- Doorway Pages: Aufbau vorgelagerter Webseiten zum "Einfangen" von Besuchern
- Hidden-Text: weisse Schrift auf weissem Grund. Fremde Schlüsselworte.
- Meta-Keywords: Nutzung fremder Schlüsselworte in den Meta-Tags.

20 Links

welche Suchmaschine füttert wen:

- <http://www.ihelpyou.com/search-engine-chart.html>,
- <http://www.zeromillion.com/webmarketing/SERP-distribution.html>

Simulator einer Suchmaschinensicht:

- <http://playground.nebel.de/sumaview/sumaview/>
- <http://www.gritechnologies.com/tools/spider.go> (Poodle Predictor)

Backlink Auswertung:

- <http://www.webuildpages.com/neat-o/>
- <http://www.seo-consulting.de/online-tools/backlink-link-check.php>

Beispiele nicht optimaler und dubioser Suchmaschinenoptimierung:

- <http://www.whirlpoolstudio-pfahler.de/> (Doubletten)
- <http://haarausfall.haarwuchs-r.de/> (Doorway)
- <http://www.holzspielzeug-discount.de/> (Doorway)
- <http://www.rabehrens.de/> (Webspam)
- <http://www.universalkat.de/> , <http://www.universalkatalysator.de/> (Cloaking)

Alternative Suchmaschinen:

- <http://www.metager.de/>
- <http://www.metager2.de/>
- <http://www.netluchs.de/>
- <http://www.yacy.net/>



21 Kontakt

Dipl.-Ing. Michael Nebel

eMail: michael@nebel.de
Telefon: +49 40 851 581 45



Freier Berater für Internet-Dienste und
Suchmaschinentechnologie



Betreiber der Suchmaschine netluchs.de



Mitglied des SuMa-eV
Gemeinnützigen Verein zur Förderung der Suchmaschinen-
Technologie und des freien Wissenszugangs



Mitarbeiter des Suchmaschinenlabors am
RRZN / [Universität Hannover](http://Universität_Hannover)
(metager.de / forschungspotal.net)