DIPL.-ING. MICHAEL JOACHIM NEBEL
NEBEL.DE

Sichere Informationsgewinnung

Ein Vortrag im Rahmen STEDAC - Fachtagung am 21.04.2006

- Das Internet gewinnt als Informationsquelle immer mehr an Bedeutung.
- Informationsentdeckung / -wiederentdeckung wird immer schwieriger.
- Suchmaschinen kommt eine zentrale Makler-Funktion zu.

=> Was sind die Folgen dieser Entwicklungen?

1 Suchverhalten im Internet

"Klassischer" Ansatz:

- Lexikon, Bibliothek, Dienstleister, Wartungsverträge
- Hochwertiger und relativ vollständiger Datenbestand
- Strukturierte Fragen
- Qualtitativ 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.

2 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.

3 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.

4 Problematische Datenquellen

Problematische Bereiche für Suchmaschinen 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)
- Breitencrawler oder Tiefencrawler

Für den nur schwer erfassbaren Teil des WWW wurde bereits 1994 von Dr. Jill Ellsworth der Begriff "deep-", "dark-" oder "invisible web" geprägt (siehe http://www.brightplanet.com/deepcontent/).

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

5 Zusätzliche Informationsquellen

Die automatisch erzeugten Daten reichen in der Regel alleine nicht aus, um die Relevanz eines Ergebnisses zu bewerten. 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
- Gramatikalische Analyse der Sprache

6 Marktanteile

	3.7.2003	7.6.2004	11.10.2005	12.04.2006	Umsatz 2004	Umsatz 2005
Google	67.1%	75.1%	83.1%	84 %	3,2 Mrd USD	6,2 Mrd USD
Yahoo	7.1%	5.8%	4.2%	4,0 %	3,5 Mrd USD (inkl Overture)	5,3 Mrd USD
MSN Web-Suche	7.7%	5.1%	4.1%	4,3 %	(2,2 Mrd USD)	(2,2 Mrd USD)
AOL Suche	1.5%	2.7%	2.5%	2,3 %	-	-
T-Online	3.4%	2.4%	1.6%	0,5 %	-	-
Altavista	1.4%	0.9%	0.4%	0,5 %	(Yahoo)	(Yahoo)
WEB.DE	1.3%	1.5%	0.5%	0,3 %	43 Mio. Euro	(United Internet: 420 Mio. Euro)
MetaGer	1.5%	1.2%	0.4%	0,4 %	-	
Lycos	2.1%	1.0%	0.4%	0,7 %	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

7 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
- Individualisierte und zielgruppenbezogene Werbung / Ergebnisseiten
- Beratung und Projekte
- Lizenzen, Produkte

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

8 Suchmaschinenoptimierung (SEO)

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

=> Die Präsentation und Postition 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"

9 Ergebnis-Manipulation

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

- Cloaking: Unterschiedliche Inhalte für Besucher und Suchmaschinen
- Link-Farm: ermogeln eines Page-Rank
- Guestbook-Spamming: ermogeln 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.

10 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

11 Zusammenfassung

- Es gibt ein Suchmaschinenmonopol.
- Es gibt ein Vermarktungsoligopol.
- Qualität der Suchergebnisse wird gezielt untergraben und weiter sinken.
  => Risiko durch falsche Informationen
- Qualitätssicherung und Bewertung der Ergebnisse muss durch den Benutzer erfolgen.
  => versteckte Kosten der Informationsgewinnung
- Treiber der aktuellen Entwickung sind die Marktführer im Suchmaschinenmarkt
  => Ausbau der Monopolstellung
- Die Markt-Monopole verhindern die Erfassung des "deep"-Web
  (Synergie von Suchmaschinenbetreibern und -optimierer)
- Der freie Zugang zu den Informationen des Internets wird schwieriger.
  (Möglichkeit der Zensur, nur wenige Suchmaschinen sind das Ziel der Spammer)

Was für Alternativen gibt es für Unternehmen / Interessensgruppen?

- Schaffung/Identifikation hochwertiger Informationsquellen
- Aufbau eigener spezialisierter Suchmaschinen

Was kann der Einzelne tun?

- Alternativen ausprobieren und Feedback geben.
- Recherchieren ist nicht nur "googlen" oder "wikipedia'n"! Fördern Sie den kritischen Umgang mit den Suchergebnissen!

12 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
Gemeinnützigenr Verein zur Förderung der Suchmaschinen-Technologie und des freien Wissenszugangs

Mitarbeiter des Suchmaschinenlabors am
RRZN / Universität Hannover
(metager.de / forschungspotal.net)