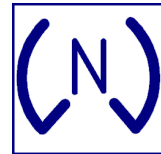


## Suchmaschinen im Internet

Ein Vortrag im Rahmen der ACM/GI Localgroup #199  
am 21. Okt 2005 in Hamburg



## Suchmaschinen im Internet

- Das Internet gewinnt als Informationsquelle immer mehr an Bedeutung
- Das Datenvolumen im Internet wächst
  
- Technische Probleme:  
Datendoubletten, nicht eindeutige Datenquellen, Werbemüll, ...
- Politische Probleme:  
Monopolisierung, Bewertung und Filterung von Inhalten, ...

=> Informationsentdeckung / -wiederentdeckung wird immer schwieriger

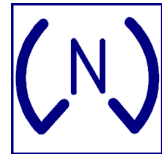
=> Suchmaschinen kommt eine zentrale Bedeutung beim Zugang zu dem im Internet gespeicherten Wissen zu

Der [Suma e.V.](#) hat sich zum Ziel gesetzt, den freien Zugang zum Wissen des Internets zu fördern.

## Informationsquellen

Woher kommt das "Wissen" des Internets?

Strukturierte Quellen	Dynamische Quellen
<ul style="list-style-type: none"><li>• Redaktionelle Inhalte, Medien</li><li>• Hersteller</li><li>• Forschungseinrichtungen</li><li>• Öffentliche Einrichtungen</li></ul>	<ul style="list-style-type: none"><li>• Newsgroups</li><li>• Mailinglisten</li><li>• Private Homepages</li><li>• Vereine</li><li>• Foren</li><li>• Blogs</li></ul>
<b>Merkmale:</b> <ul style="list-style-type: none"><li>- bekannte Quellen</li><li>- häufig für Suchmaschinen optimiert</li><li>- zu weiten Teilen von Suchmaschinen erfasst</li></ul>	<b>Merkmale:</b> <ul style="list-style-type: none"><li>- entstehen und vergehen ungeordnet</li><li>- häufig "subjektive" aber auch hochwertige Inhalte</li><li>- schwierig zu erfassen</li></ul>



## 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

## Deep-Web

"deep-", "dark-" oder "invisible web" - kurz was Suchmaschinen nicht sehen.

- Suchmaschinen folgen Links von Seite zu Seite
- oftmals nur die "obersten" Ebenen indiziert (Surface Web)
- Datenbanken, Foren usw. nur teilweise erfasst (Eingabefelder, "Crawlerfallen")
- Erfasst deutschsprachige Seite der grossen Suchmaschinen ~100 Mio. Seiten (geschätzt!)
- Umfang des deutschsprachigen Netzes: über 1 Mrd. Seiten (geschätzt)
- Inhalte oftmals sehr hochwertig

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

Der Begriff wurde schon 1994 von Dr. Jill Ellsworth geprägt. (siehe <http://www.brightplanet.com/deepcontent/>)

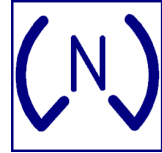
## Suchmaschinenoptimierung

- 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"



## 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

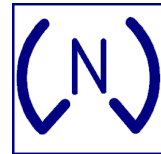
- Webspammer bauen automatisiert Internetseiten (Webspam-Seiten)
  - Die Webspam-Seiten sind mit Anzeigen versehen
  - Üblicherweise werden alte aber etablierte Domains für solche Angebote aufgekauft
  - Suchmaschinen erfassen diese Seiten und geben sie als Ergebnisse mit aus
  - Ein kleiner Prozentsatz 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

## Geschäftsmodelle für Suchmaschinen

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

- 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.



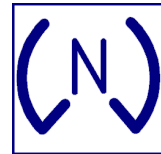
## Marktanteile

	3.7.2003	7.6.2004	18.5.2005	11.10.2005	Umsatz 2004
Google	67.1%	75.1%	81.7%	83.1%	3,2 Mrd USD
Yahoo	7.1%	5.8%	3.8%	4.2%	3,5 Mrd USD (inkl Overture)
MSN Web-Suche	7.7%	5.1%	4.7%	4.1%	-
AOL Suche	1.5%	2.7%	2.5%	2.5%	-
T-Online	3.4%	2.4%	1.8%	1.6%	-
Altavista	1.4%	0.9%	0.5%	0.4%	(Yahoo)
WEB.DE	1.3%	1.5%	0.9%	0.5%	43 Mio. Euro
MetaGer	1.5%	1.2%	0.6%	0.4%	-
Lycos	2.1%	1.0%	0.4%	0.4%	103 Mio. Euro (Europa)

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

## Suchmaschineninfrastruktur

Komponente	Beschreibung	Ressourcen
Crawler/Fetcher	Sammeln der Webseiten	Netzwerk-Bandbreite
Parser/Indexer	Analyse der Dokumente	CPU / RAM
Datenspeicher	Indices und gecachte Dokumente	Festplattenspeicher
Suchfrontend	Interface für die Benutzer	CPU-Last, I/O
Monitoring	Überwachung der Komponenten	Schnittstellen, Aggregationen
Wartung/Betrieb	Hardware, Software	Anzahl der Systeme



## Eckdaten einer Suchmaschine

**Ziel:** eine Suchmaschine 100 Mio. indizierten Dokumenten und 5 Mio Suchanfragen / Monat

### Festplattenplatz:

- je Dokument ~ 25 kB + 5 kB Metainformationen => 3 TB Festplattenspeicher

### Netzwerk Crawler:

- alle drei Monate vollständige Aktualisierung aller Seiten
- Netzwerk Overhead: 20% => 1,2 TB / Monat
- 1 Tag = 16 Stunden
- 35 Mio. Dokumente / Monat => 20 Dokumente / Sekunde crawlen => 4 Mb/s Bandbreite

### Frontend:

- 5 Mio. Suchanfragen / Monat, 1 Tag = 12 Stunden => 4 Anfragen / Sekunde

## Freie Suchmaschinensoftware

<a href="#">ht://Dig</a>	Einfache Installation	~ 50.000 Seiten
<a href="#">Harvest</a>	Feldbasierte Suche, sehr viele Möglichkeiten	~ 200.000 Seiten
<a href="#">mnoGoSearch</a>	PHP Erweiterung, einfache Installation	~ 300.000 Seiten
<a href="#">ASPseek</a>	C++ geschrieben, keine Weiterentwicklung	~3.000.000 Seiten
<a href="#">Nutch</a>	Java, sehr aktive Entwicklung	>> 10.000.000 Seiten
<a href="#">Heritrix</a>	reiner Crawler, kein Frontend, Lucene Datenbank	
<a href="#">Terrier</a>	Framework zur Suchmaschinenerstellung	
<a href="#">Lucene</a>	Bibliothek zur Volltextsuche (Java)	

## Bessere Suchmaschinen?

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

Problem: Doppelbedeutungen von Suchwörtern

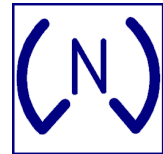
Ziel: automatische Hinführung zur "subjektiv richtigen" Begriffswelt

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!



## Von Google lernen

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

## Beispiel: netluchs.de

### Ziel:

Aufbau einer deutschsprachigen Internetsuchmaschine mit mindestens 50 Mio. Dokumenten.

### Ansatz:

Einsatz von Nutch, Erweiterung der Architektur ([Blockmodell](#))

### Status:

- Aufbau des Prototypen abgeschlossen: ~ 6 Millionen indizierte Dokumente.
- Zulieferung zu [Metager.de](#)

### Erfahrungen:

- komplexe Lösung
- sehr gute Skalierung und Verteilung
- kritische Parameter: Festplattenplatz und Geschwindigkeit
- Monitoring aggregierter Ziffern essentiell
- Endausbau: 10 - 20 Server ([Systemskizze](#))
- sehr aktive Weiterentwicklung der Software (<http://lucene.apache.org/nutch/>)



## Beispiel: metager2.de

### Ziel:

Fortführung von metager.de. Erweiterung um Spamsicherheit.

### Ansatz:

Nach der Zusammenfassung der Ergebnisse der nachgelagerten Suchmaschinen werden die Ergebnisse nachgeladen und geprüft. ([Blockmodell](#))

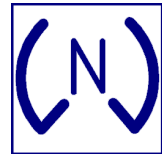
### Status:

- Erfolgreich am Netz
- Erweiterung um "Fist-Level-Suchmaschine"

### Erfahrungen:

- benötigt recht viele Ressourcen (CPU, Netz)
- skaliert über mehrere Maschinen

**Metager<sup>2</sup>**



## Beispiel: Yacy

### Ziel:

Nicht monopolisierbare Suchmaschine

### Ansatz:

Peer-To-Peer-Suchmaschine

### Status:

- Software robust und frei verfügbar
- "Scraping"-Proxy: erfasst Seiten während des Surfens
- Möglichkeiten der Publikation von Informationen im yacy-Netzwerk
- Daten werden redundant im Netzwerk gehalten

### Erfahrungen:

- findet Inhalte ausserhalb der "Main-Stream"-Seiten
- Suchdauer vergleichbar mit einer Metasuche



## Konzept: Minisucher

### Ziel:

Nicht monopolisierbare dezentrale Suchmaschine und Erfassung des Deep-Web

### Ansatz:

Hierarchisches Konzept basierend auf vielen kleinen Minisuchern. Jeder Minisucher ist für die Qualität seiner Inhalte verantwortlich. ([Blockmodell](#))

### Status:

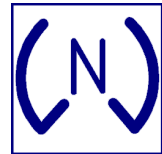
- Konzept

## Zusammenfassung

- Es gibt ein Suchmaschinenmonopol
- Es gibt ein Vermarktungsoligopol
- Treiber der aktuellen Entwicklung 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 kann der Einzelne tun?

- 1.) Alternativen ausprobieren und Feedback geben.
- 2.) Recherchieren ist nicht nur "googlen" oder "wikipedia'n" ! Fördern Sie den kritischen Umgang mit den Suchergebnissen!
- 3.) Dem Suma e.V. helfen



## Kontakt

Dipl.-Ing. Michael Nebel  
michael@nebel.de

Internet:  
<http://www.nebel.de/>  
<http://www.netluchs.de/>