

Internetsuchmaschinen

- Kritische Betrachtung von Funktionalität und Leistungsfähigkeit im Vergleich

Prof. Dr. Dirk Lewandowski

Hochschule für Angewandte Wissenschaften Hamburg, Fakultät DMI, Department Information

Programm heute

- **14.00 - 15.30 Uhr: Internetsuchmaschinen - Kritische Betrachtung von Funktionalität und Leistungsfähigkeit im Vergleich (Vortrag)**
 - Vergleich der Leistungsfähigkeit (Qualität) der bekannten Suchmaschinen
 - kurze Pause
 - Überblick Recherchemöglichkeiten
- **16.00 - 17.30 Uhr: Gruppenweise Durchführung von Rechercheprojekten**
 - Recherchieren in der Gruppe
 - Vorstellung der Recherchen im Plenum

Gliederung

Welche Suchmaschinen sind wichtig?

Wie Suchen Nutzer im Web?

Wie kann man die Qualität von Suchmaschinen messen?

Qualität des Index

Qualität der Treffer

Recherchemöglichkeiten

Gliederung

Welche Suchmaschinen sind wichtig?

Wie Suchen Nutzer im Web?

Wie kann man die Qualität von Suchmaschinen messen?

Qualität des Index

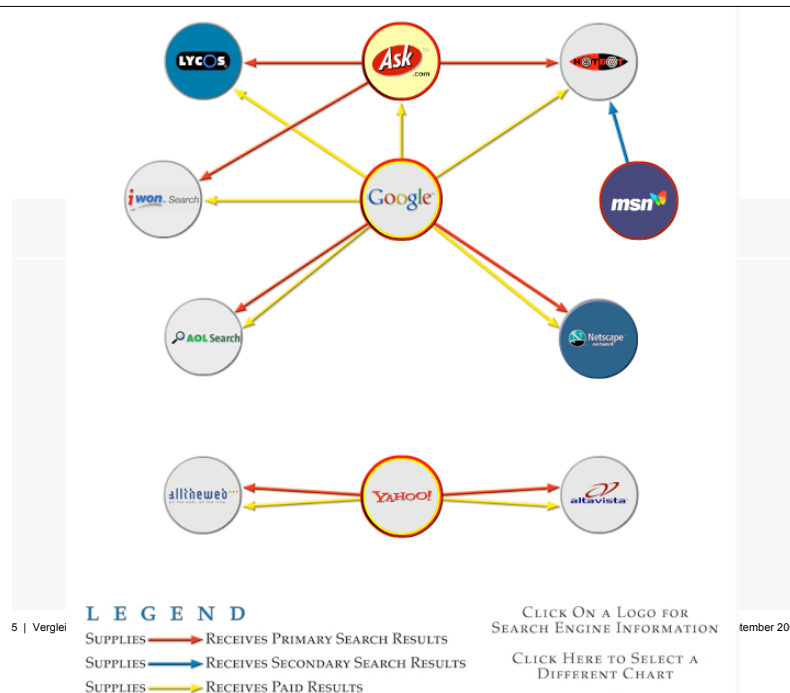
Qualität der Treffer

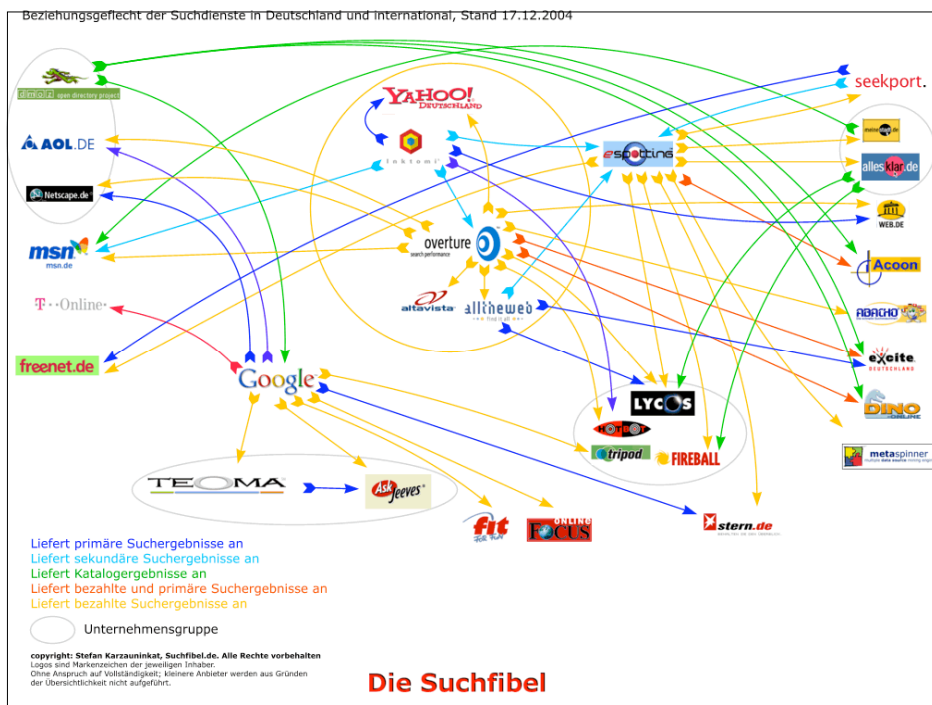
Recherchemöglichkeiten

Welche Suchmaschine soll ich verwenden?

4 | Vergleich von Internetsuchmaschinen - Dirk Lewandowski

17. September 2007





Gliederung

Welche Suchmaschinen sind wichtig?

Wie Suchen Nutzer im Web?

Wie kann man die Qualität von Suchmaschinen messen?

Qualität des Index

Qualität der Treffer

Recherchemöglichkeiten

Grunddaten zum Nutzerverhalten

- **Boolesche Operatoren**

- Bei weniger als 10% der Anfragen; keine Veränderung im Lauf der Jahre (Spink&Jansen 2004).
- Nur etwa die Hälfte der Nutzer weiß, dass es Operatoren gibt (Machill et al. 2003).
- 20% der Nutzer geben an, sie öfter zu verwenden (Machill et al. 2003).
- Etwa die Hälfte der Anfragen mit Operatoren enthält Fehler (Spink et al. 2000), bei der Verwendung von Plus-/Minuszeichen lag die Fehlerquote bei 2/3.

- **Profisuche**

- 59% der Nutzer kennen die Profisuche (Machill et al. 2003).
- Nur 14% nutzen sie öfter (Machill et al. 2003).

- **Auswertung der Treffer**

- 80% der Nutzer betrachten nur die erste Ergebnissseite (= 10 Treffer) (Spink&Jansen 2004).
- Vor allem die Treffer auf den ersten Listenplätzen werden angesehen; möglichst kein Scrollen
- Pro Session werden nur etwa fünf Dokumente gesichtet (Spink&Jansen 2004).
- Sessions dauern weniger als 15 Minuten.

8 | Vergleich von Internetsuchmaschinen - Dirk Lewandowski

17. September 2007

Drei Anfragetypen

Anfragetypen nach Broder (2002)

- **Informational (informationsorientiert)**

- Nutzer möchte sich zu einem Thema informieren.
- Ziel sind mehrere Dokumente.

- **Navigational (navigationsorientiert)**

- Ziel ist es, eine bestimmte Seite (wieder) zu finden.
- Typisch: Suche nach Homepage („Daimler Chrysler“).
- Ziel ist i.d.R. ein Dokument.

- **Transactional (transaktionsorientiert)**

- Ziel ist das Auffinden einer Website, auf der dann eine Transaktion stattfinden soll.
- Beispiele für Transaktionen: Kauf eines Produkts, Download einer Datei.

9 | Vergleich von Internetsuchmaschinen - Dirk Lewandowski

17. September 2007

Gliederung

Welche Suchmaschinen sind wichtig?

Wie Suchen Nutzer im Web?

Wie kann man die Qualität von Suchmaschinen messen?

Qualität des Index

Qualität der Treffer

Recherchemöglichkeiten

Die Qualität von Suchmaschinen kann nur durch eine Kombination unterschiedlicher Faktoren gemessen werden.

Qualitätsfaktoren für Suchmaschinen

- **Qualität des Index**
 - Größe des Datenbestands, Abdeckung des Web.
 - Abdeckung bestimmter Bereiche (Sprachräume, Länder).
 - Überschneidungen der Indices.
 - Aktualität des Datenbestands.
- **Qualität der Suchresultate**
 - Retrievaleffektivität
 - Zufriedenheit der Nutzer
 - Überschneidungen der (Top-)Ergebnisse.
- **Qualität der Suchfunktionen**
- **Nutzerfreundlichkeit + Benutzerführung**

Gliederung

Welche Suchmaschinen sind wichtig?

Wie Suchen Nutzer im Web?

Wie kann man die Qualität von Suchmaschinen messen?

Qualität des Index

Qualität der Treffer

Recherchemöglichkeiten

Web-Abdeckung

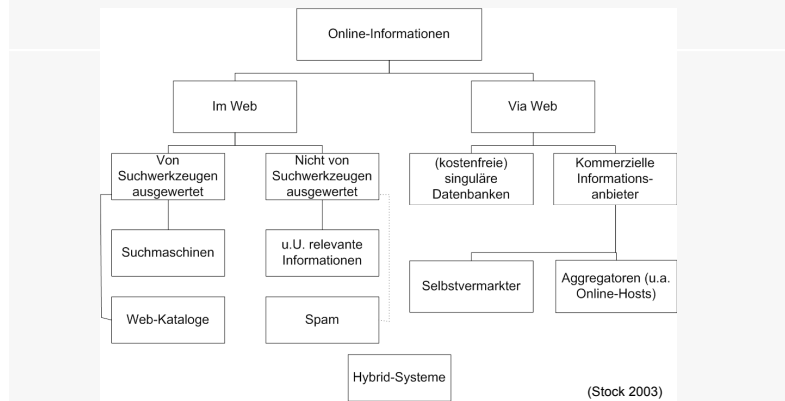
Keine Suchmaschine deckt das gesamte Web ab!

- Aber: keine genauen Zahlen; aktuelle Untersuchungen fehlen.
- Größe des Web ist unbekannt; die großen Suchmaschinen haben einige Milliarden Dokumente indexiert.
- Wenig erforschter Bereich: Invisible Web.

Definitionen Invisible/Deep Web

- “Text pages, files, or other often high-quality authoritative **information** available via the World Wide Web **that general-purpose search engines cannot**, due to technical limitations, or will not, due to deliberate choice, **add to their indices** of Web pages” (Sherman u. Price 2001).
- “The deep Web - those pages do not exist until they are created dynamically as the result of a specific search“ (Bergman 2001).

Taxonomie der digitalen Online-Information



14 | Vergleich von Internetsuchmaschinen - Dirk Lewandowski

17. September 2007

Country Bias

Tabelle 3.1. Prozentsatz der abgedeckten Websites (Vaughan u. Thelwall 2004, 700)

	U.S.	China	Singapore	Taiwan	Durchschnitt
Google	87%	70%	56%	75%	72%
AllTheWeb	83%	61%	50%	75%	67%
AltaVista	80%	52%	41%	4%	44%
Average	83%	61%	49%	51%	61%

15 | Vergleich von Internetsuchmaschinen - Dirk Lewandowski

17. September 2007

Überschneidung der Datenbestände

- **Überschneidung der Indices**
 - Relativ gering, daher lohnt sich die Suche in mehreren Suchmaschinen.
- **Überschneidung der Suchergebnisse**
 - Sehr gering innerhalb der Top10,20.

Überschneidung der Datenbestände

Table 3
Search result overlap on the first results page

	Unique	Two engines	Three engines	All four engines
Google only	94,293			
Yahoo! only	106,057			
Ask Jeeves only	115,525			
MSN Search only	96,371			
Google & Yahoo!		7175		
Google & Ask Jeeves		17,279		
Google & MSN Search		7824		
Yahoo! & Ask Jeeves		5519		
MSN Search & Yahoo!		14,039		
MSN Search & Ask Jeeves		3679		
MSN Search & Google		5336		
Google, Yahoo!, & Ask Jeeves			4002	
Google, Yahoo!, & MSN Search			3713	
Yahoo!, Ask Jeeves & MSN Search			2510	
Google, Ask Jeeves & MSN Search			2173	
Yahoo!, Google, MSN Search & Ask Jeeves				5301
Total = 485,460	412,246 (84.9%)	55,515 (11.4%)	12,398 (2.6%)	5301 (1.1%)

Aktualität der Datenbestände

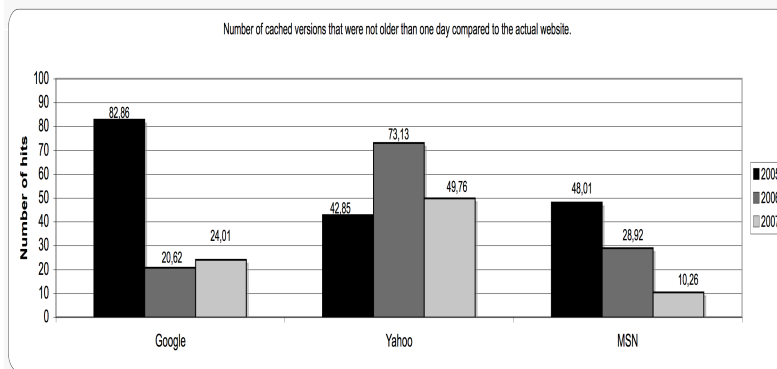
- **Wichtig wegen**

- Auffinden und Erfassen neuer Dokumente.
- Aktualisierung veränderter Dokumente; korrekte Repräsentation der URLs.

- **Untersuchung 2005-2007**

- Beobachtungszeitraum jeweils sechs Wochen.
- 40 täglich aktualisierte Seiten (+ 30 unregelmäßig aktualisierte).

Auch aktuelle Seiten werden nicht zuverlässig regelmäßig indexiert.



Aktualität der Datenbestände

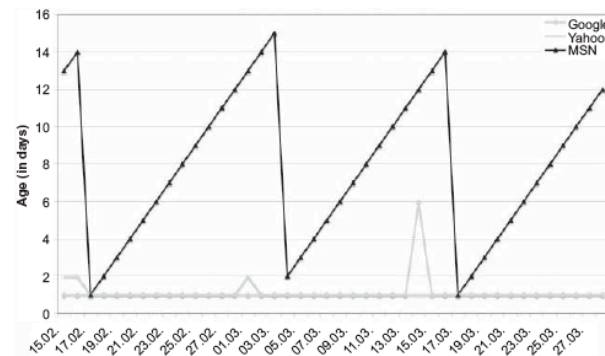


Fig. 13. Typical MSN update pattern (up-to-dateness for <http://www.seesener-beobachter.de>).
 20 | Vergleich von Internetsuchmaschinen - Dirk Lewandowski 17. September 2007
 Quelle: Lewandowski et al. 2006

Aktualität der Datenbestände

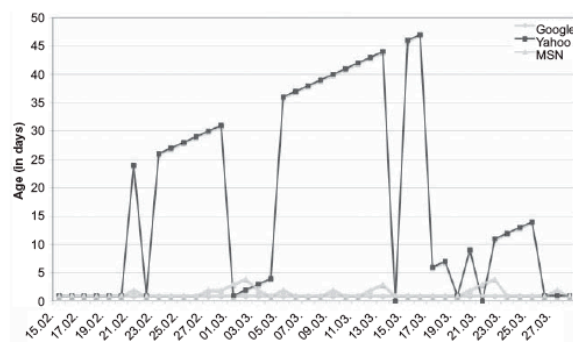
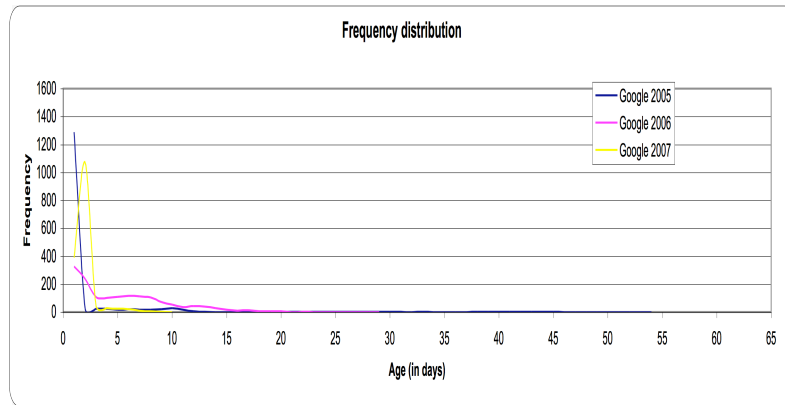


Fig. 14. Yahoo up-to-dateness for <http://www.sportschau.de>.
 21 | Vergleich von Internetsuchmaschinen - Dirk Lewandowski 17. September 2007
 Quelle: Lewandowski et al. 2006

Aktualität der Datenbestände



22 | Vergleich von Internetsuchmaschinen - Dirk Lewandowski

17. September 2007

Aktualität der Datenbestände

Table 2 Oldest and newest pages found by the search engines

Search Engine	Newest Page Found		Oldest page found	
	Daily Updated Pages	Control Group	Daily Updated Pages	Control Group
2005				
Google	0	-	54	-
Yahoo	0	-	62	-
MSN	0	-	17	-
2006				
Google	0	0	29	253
Yahoo	0	-	32	-
MSN	0	0	16	19
2007				
Google	0	1	10	175
Yahoo	0	-	26	-
MSN	0	1	30	30

23 | Vergleich von Internetsuchmaschinen - Dirk Lewandowski

17. September 2007

Aktualität der Datenbestände

Table 1 Average values for all pages that are updated on a daily basis

Search Engine	2005		2006		2007	
	Mean	Median	Mean	Median	Mean	Median
Google	3.1	1	5.6	5	2.2	2
Yahoo	9.8	4	2.3	1	2.7	2.5
MSN/Live.com	3.5	1	3.3	2	5.7	3

Table 3 Average values for all pages that are updated irregularly

Search Engine	2006		2007	
	Mean	Median	Mean	Median
Google	20.6	13.5	14.8	6
MSN/Live.com	7.7	8	9.3	9

Gliederung

Welche Suchmaschinen sind wichtig?

Wie Suchen Nutzer im Web?

Wie kann man die Qualität von Suchmaschinen messen?

Qualität des Index

Qualität der Treffer

Recherchemöglichkeiten

Retrievaleffektivität

Klassische Tests

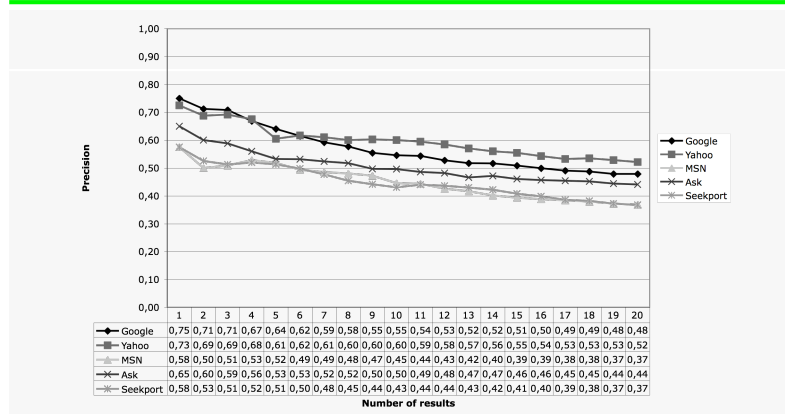
- **Welche Suchmaschine liefert die besten Treffer?**
- **Test anhand von**
 - einer bestimmten Anzahl von Suchanfragen (50).
 - einer bestimmten Anzahl von Treffern pro Suchmaschine (20).
 - Bewertungen durch Juroren (wichtige Frage: *wer* bewertet?).
 - Ja/Nein-Entscheidungen über Relevanz, verschiedenen Skalen.

Precision ist das am häufigsten verwendete Retrievalmaß.

Klassische Retrievalmaße

- **Precision**
 - Anteil der relevanten ausgegebenen Treffer an der Gesamtheit der ausgegebenen Treffer.
 - Wird bei Suchmaschinen i.d.R. bis zu einem Cut-Off-Wert bestimmt (z.B. 20).
- **Recall**
 - Anteil der relevanten ausgegebenen Treffer an der Gesamtheit aller überhaupt vorhandenen Treffer (= aller im Web vorhandener relevanter Dokumente).
 - Bei Suchmaschinen nicht messbar (Ausweg evtl. über Pooling-Methoden).
- **Fallout**
 - Anteil der ausgegebenen nicht relevanten Treffer an der Gesamtzahl der nicht relevanten Treffer im Datenbestand.
- **Generality**
 - Anteil der relevanten Dokumente im zugrunde liegenden Datenbestand.

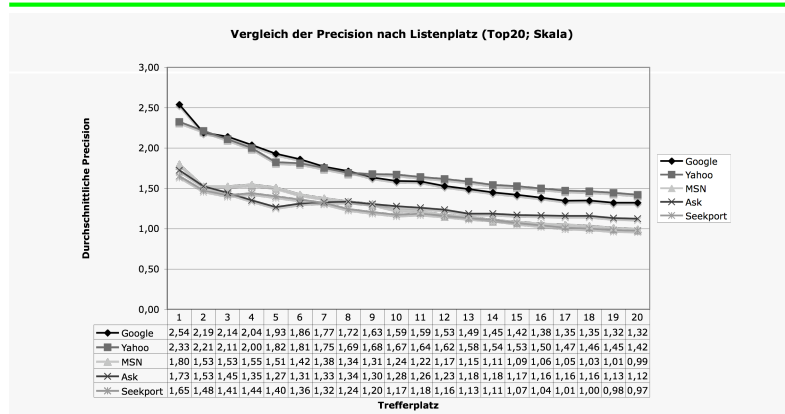
Precision@10 zwischen 0,43 und 0,60.



28 | Vergleich von Internetsuchmaschinen - Dirk Lewandowski

17. September 2007

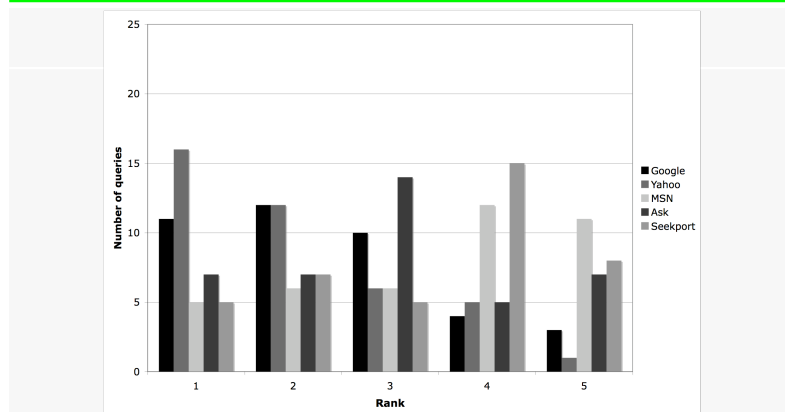
Skalenbewertung zeigt die Gruppierung der Suchmaschinen noch deutlicher.



29 | Vergleich von Internetsuchmaschinen - Dirk Lewandowski

17. September 2007

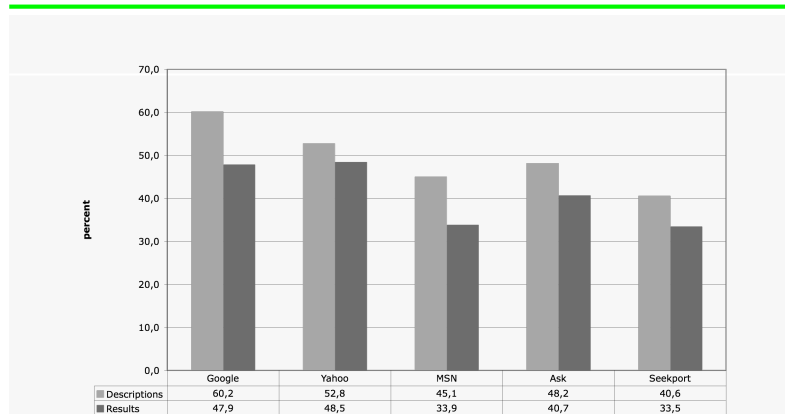
Keine klare Reihung der Suchmaschinen, wenn einzelne Anfragen betrachtet werden.



30 | Vergleich von Internetsuchmaschinen - Dirk Lewandowski

17. September 2007

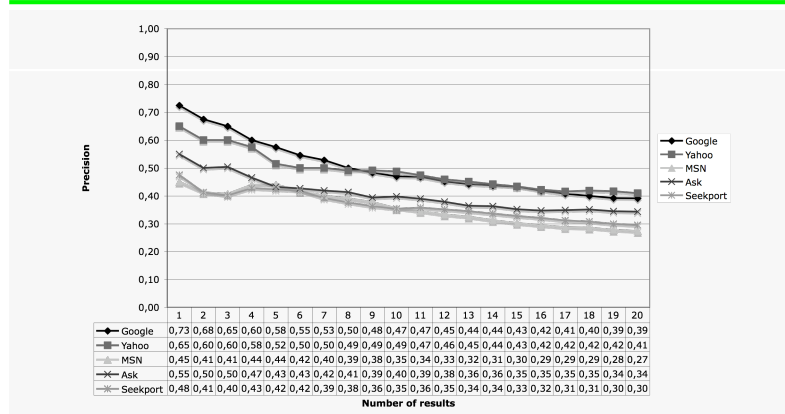
Bei allen Suchmaschinen mehr relevante Beschreibungen als relevante Treffer.



31 | Vergleich von Internetsuchmaschinen - Dirk Lewandowski

17. September 2007

„Super-Precision“ (Beschreibungen und Treffer relevant).



32 | Vergleich von Internetsuchmaschinen - Dirk Lewandowski

17. September 2007

Offene Frage: Wie können Retrievalmaße dem spezifischen Verhalten der Suchmaschinen-Nutzer angepasst werden?

Problembereiche

- **Such-Sessions**
 - Nur Treffer auf der ersten Ergebnisseite (bzw. auf dem direkt sichtbaren Bildschirmbereich) werden angesehen.
 - In der Regel werden maximal fünf Dokumente angesehen.
 - Recherche wird oft abgebrochen, sobald ein passendes Dokument gefunden wurde.
 - Wie können die Maße dem Bedürfnis nach nur einem/wenigen relevanten Treffern angepasst werden?
- **Zusammensetzung der Trefferlisten**
 - Abbildung mehrerer Facetten eines Themas auf den vorderen Plätzen der Trefferlisten.
 - Wie kann in der Evaluierung auf Facetten eingegangen werden?
- **Interaktion**
 - Allgemeines IR-Evaluationsproblem: Recherche ist immer als interaktiver Prozess anzusehen.
 - Wie können die Maße an den tatsächlichen Suchprozess angepasst werden?

33 | Vergleich von Internetsuchmaschinen - Dirk Lewandowski

17. September 2007

Beispiel einer Trefferliste mit mehreren Facetten (mehrdeutige Anfrage „Jaguar“)

Jaguar DE - Jaguar Cars
Der Jaguar XJ aus Aluminium. Der Jaguar XJ aus Aluminium. Auch als Cabriolet erhältlich. Auch als Cabriolet erhältlich ...
www.jaguar.com/de/ - 21k - [Im Cache](#) - [Ähnliche Seiten](#)

Jaguar - [Diese Seite übersetzen]
Official worldwide web site of Jaguar Cars.
www.jaguar.com/ - [Ähnliche Seiten](#)

Jaguar Panthera onca
Nachschlagewerk zu den Wildtieren dieser Welt. - Der Jaguar.
www.tierencyklopaedie.de/tiere/jaguar.html - 9k - [Im Cache](#) - [Ähnliche Seiten](#)

Jaguar Cars
www.jaguarcars.com/de/ - 1k - [Im Cache](#) - [Ähnliche Seiten](#)

Jaguar - Wikipedia
eine Raubkatze, siehe **Jaguar** (Katze); einen britischen Automobilhersteller, siehe **Jaguar** (Auto); das Formel 1-Team dieses Automobilherstellers, ...
[de.wikipedia.org/wiki/Jaguar_\(Katze\)](http://de.wikipedia.org/wiki/Jaguar_(Katze)) - 32k - 22. Juni 2006 - [Im Cache](#) - [Ähnliche Seiten](#)
[Weitere Ergebnisse von de.wikipedia.org]

Jaguar (Katze) - Wikipedia
Bei vielen indischen Völkern hatte oder hat der **Jaguar** eine bedeutende Rolle in ...
- elegante Raubkatze des Dschungels Webseite für den **Jaguar** ...
[de.wikipedia.org/wiki/Jaguar_\(Katze\)](http://de.wikipedia.org/wiki/Jaguar_(Katze)) - 32k - 22. Juni 2006 - [Im Cache](#) - [Ähnliche Seiten](#)
[Weitere Ergebnisse von de.wikipedia.org]

Jaguar Stahlwarenfabrik GmbH & Co. KG
Die Solinger Firma **Jaguar** Stahlwarenfabrik GmbH & Co. KG ist Hersteller professioneller Haarscheren für den Friseurbedarf.
www.jaguar-solingen.com/ - 7k - [Im Cache](#) - [Ähnliche Seiten](#)

JAG - Jaguar Association Germany - Willkommen - Welcome
Jaguar Association Germany (JAG). Offizielle Homepage des größten deutschen Jaguar Oldtimer Clubs, **Jaguar** Veranstaltungstermine, **Jaguar** Werkstatt, ...
www.jaguar-association.de/ - 14k - [Im Cache](#) - [Ähnliche Seiten](#)

Amazon.de: Jaguar: DVD
Jaguar, Jean Reno, Patrick Bruel, Francis Veber, Hanissou Lowe, Vladimir Cosma.
www.amazon.de/exec/obidos/ASIN/B000917KW4 - 44k - [Im Cache](#) - [Ähnliche Seiten](#)

17. September 2007

Gliederung

Welche Suchmaschinen sind wichtig?

Wie Suchen Nutzer im Web?

Wie kann man die Qualität von Suchmaschinen messen?

Qualität des Index

Qualität der Treffer

Recherchemöglichkeiten

Grundlegende Suchfunktionen

- **Boolesche Operatoren**
 - Yahoo, MSN, Ask: vollständige Unterstützung.
 - Google: keine vollständige Unterstützung, OR arbeitet unzuverlässig, Klammerung funktioniert nicht.
- **Phrasensuche**
 - funktioniert bei allen
- **Abstandsoperatoren**
 - nur bei Exalead

Suche über Felder

in der erweiterten Suche oder über Befehle

- **Titel:**
 - *title:*
- **Dateityp:**
 - *filetype:*
- **URL:**
 - *inurl:*
- **bestimmter Server:**
 - *site:*
- **Land (über Top-Level-Domain)**
 - *site:*
- **Verlinkung:**
 - *link:*
- **Sprache**
- **Datum**

Funktionsfähigkeit der Datumsbeschränkung

Table III Error ratio in the search engines analysed

Search engine	Number of documents classified correctly	Number of documents classified incorrectly	Error ratio (%)
Teoma	115	193	62.66
Google	166	113	40.50
Yahoo!	154	130	45.77

Table IV Improvements and deteriorations caused by the use of date-restricted queries

Search engine	Deteriorations	No alteration	Improvements
Teoma	14	17	16
Google	8	12	25
Yahoo!	7	10	30

Weitere Suchmöglichkeiten

- **Ähnliche Seiten**
- **Cache Copies**
 - Google, Yahoo, MSN
 - bei Ask teilweise
- **Alte Versionen von Websites**
 - Wayback Machine: www.archive.org

Benutzerführung: Einschränkung der Suchergebnisse

- **Themenbeschränkung**
 - www.seekport.de
- **Clustering**
 - www.clusty.com (Datenbasis: Metasuche)
 - www.web.de (Datenbasis: Yahoo*)
- **Visualisierung von Suchergebnissen**
 - www.kartoo.com (Datenbasis: Metasuche)
- **Vorschläge für Suchanfragen**
 - www.yahoo.de
 - www.ask.de
- **Manuelle Veränderung des Rankings**
 - mindset.research.yahoo.com
 - www.msn.de

40 | Vergleich von Internetsuchmaschinen - Dirk Lewandowski

17. September 2007

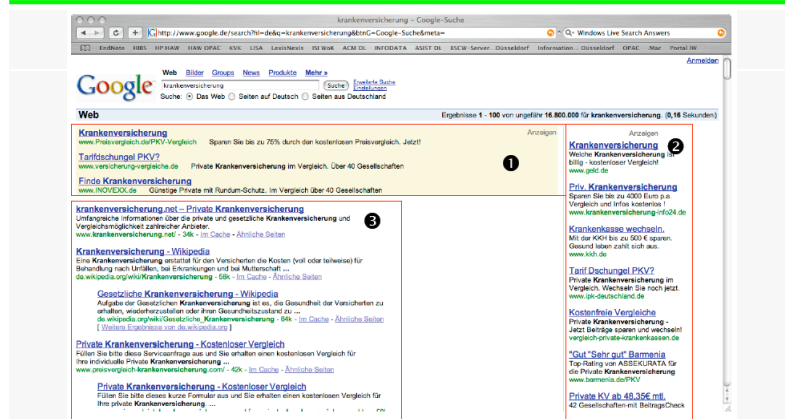
Dokumentkollektionen

Tabelle 2 Kollektionen bei den verschiedenen Suchmaschinen

Kollektion	Google	Yahoo	MSN	Ask
News	x	x	x	
Bilder	x	x	x	x
Wissenschaft	x		x	
Lokale Suche	x	x		
Produkte	x	x	x	x
Software				
Blogs	x		x	x
Bücher	x			
Verzeichnis	x	x		
Video	x	x		
Audio		x		
Frage-Antwort-Dienst		x		
Enzyklopädie			x	

17. September 2007

Präsentation der Trefferlisten



17. September 2007

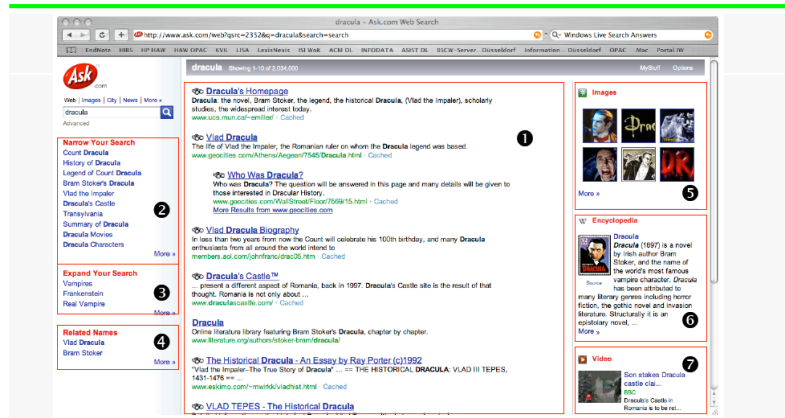
Präsentation der Trefferlisten

<http://www.google.com/search?hl=en&q=nosferatu&btnG=Search>

43 | Vergleich von Internetsuchmaschinen - Dirk Lewandowski

17. September 2007

Präsentation der Trefferlisten



44 | Vergleich von Internetsuchmaschinen - Dirk Lewandowski

17. September 2007

Vielen Dank für Ihre
Aufmerksamkeit.

Prof. Dr. Dirk Lewandowski
www.durchdenken.de/lewandowski
E-Mail: dirk.lewandowski@haw-hamburg.de