

openHPI: **Wie funktioniert eine Suchmaschine?**
Websuche

Dr. Ralf Krestel
Hasso-Plattner-Institut
Universität Potsdam

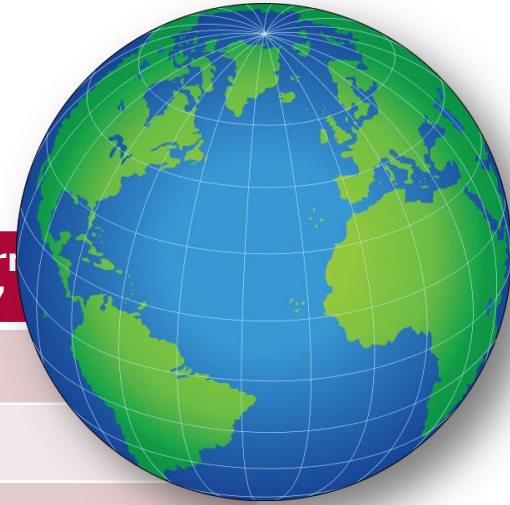
- Heterogenität
 - Viele unterschiedliche Benutzer, Themen, Sprachen, Dokumenttypen,...
 - Webseiten sind nicht wie gedruckte Dokumente (dynamischer Inhalt,...)
 - Offen für Alle: Meinungen, Schreibstile,...
- Hyperlinks
 - Dokumente sind mit einander verbunden und beziehen sich aufeinander
- Problemgröße
 - Sehr viele Dokumente, Anfragen, schnell sich ändernder Inhalt,...
- Spam
 - „Das Böse ist immer und überall“ (EVA 1985)
- Geschäftsmodell
 - Websuchmaschine zu betreuen/unterhalten ist sehr teuer...

Websuche

Heterogenität der Nutzer

■ Internetnutzung in Millionen

Region	Bevölkerung 2000	Bevölkerung 2017	Internetnutzer 2000	Internetnutzer 2017
Afrika	767	1246	5	335
Asien	3557	4148	114	1856
Europa	729	822	105	630
Mittlerer Osten	176	250	3	141
Nordamerika	307	363	108	320
Latinamerika	593	647	18	384
Ozeanien	30	40	8	27
Welt	6007	7018	361	2406



Websuche

Heterogenität der Sprachen

- Schätzung der Sprache der Top-10 Millionen Webseiten und Anzahl Wikipediaseiten

Rang	Webseiten (2017)	Prozent	Wikipediaseiten (2017)
1	Englisch	52.1%	2,567,509
2	Russisch	6.5%	318,850
3	Japanisch	5.6%	523,639
4	Deutsch	5.5%	808,044
5	Spanisch	5.1%	402,430
6	Französisch	4.1%	709,312
7	Portugisisch	2.6%	429,730
8	Italienisch	2.3%	499,234
9	Chinesisch	2.0%	205,047
10	Polnisch	1.7%	539,688



https://commons.wikimedia.org/wiki/File:Globe_Atlantic.svg
https://w3techs.com/technologies/overview/content_language/all
https://en.wikipedia.org/wiki/Wikipedia:Multilingual_statistics

- Einige der Dokumentarten, die eine Suchmaschine verarbeiten sollen könnte:
 - application/ms-excel (different versions),
application/mspowerpoint (different versions),
application/msword (different versions),
application/pdf (different versions), application/postscript,
application/x-dvi, application/x-tar, application/x-zip-compressed, text/html (different versions and encodings),
text/plain (different encodings), text/rtf, application/xml,
text/xml, application/xhtml+xml, application/docbook+xml,
application/x-shockwave-flash, ...

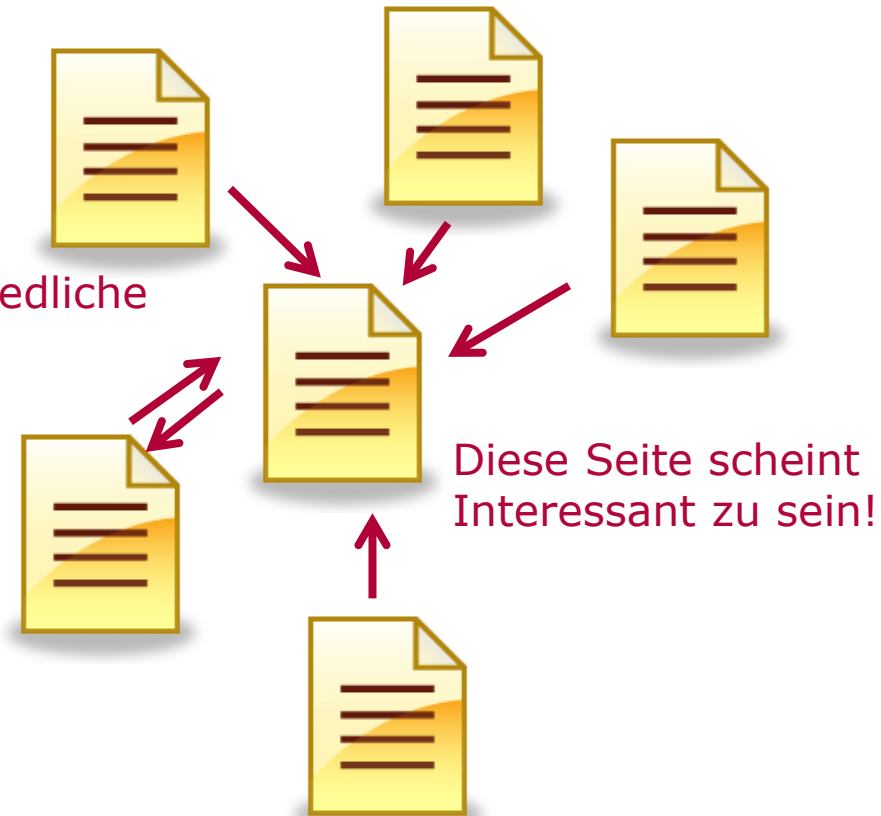
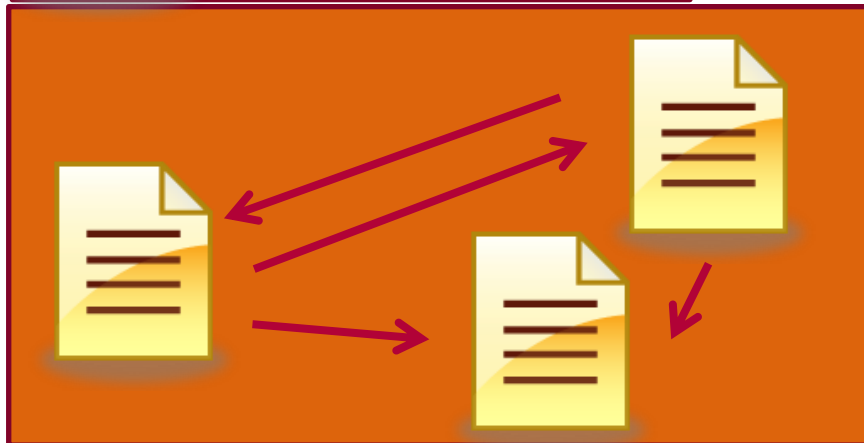
- Wie sieht es mit
 - Bildern, Videos, Audiodateien, ausführbarem Code, ... aus?

- Websuchmaschinen werden für unterschiedliche Aufgaben verwendet
- Drei Hauptarten von Anfragen (Broder 2002):
 - **Informationsorientiert** (~50%)
 - „Relativitätstheorie“, „Hauptstadt von Taiwan“,...
 - **Navigationsorientiert** (~20%)
 - „facebook“, „amazon“, „ebay“,...
 - **Transaktionsorientiert** (~30%)
 - „hotels hamburg“, „flug berlin paris“, „eminem mp3“, skype download“,...
 - (Verknüpfungsorientiert)
 - „link:www.hpi.uni-potsdam.de“,...

- Dokumente im Web können sich gegenseitig **verlinken**
- Links werden **nicht zufällig** gesetzt



Unterschiedliche
Themen?



Websuche

Größe: Anzahl der Anfragen

Jahr	Anzahl von Google-Suchen pro Jahr	Durchschnittliche Anzahl von Suchen pro Tag
2015	2,834,650,000,000	7,766,000,000
2014	2,095,100,000,000	5,740,000,000
2013	2,161,530,000,000	5,922,000,000
2012	1,873,910,000,000	5,134,000,000
2011	1,722,071,000,000	4,717,000,000
2010	1,324,670,000,000	3,627,000,000
2009	953,700,000,000	2,610,000,000
2008	637,200,000,000	1,745,000,000
2007	438,000,000,000	1,200,000,000
2000	22,000,000,000	60,000,000
1998	3,600,000	9,800

~90.000 pro
Sekunde

- Schätzung oder Betreiberangaben
- Wie schätzt man Indexgröße?
 - Ausnutzen des Zipfschen Gesetzes
 - 50 zufällige Wörter als Anfrage
 - Hochrechnung über Trefferzahl

Suchmaschine (Jahr) Schätzung	Anzahl indexierter Webseiten
Google (2016)	~47,000,000,000,000
Bing (2016)	~17,000,000,000,000
Google (2014)	~15,000,000,000,000
Bing (2014)	~9,000,000,000,000
Yahoo (2010)	~50,000,000,000
Ask (2010)	~1,700,000,000

Jahr (Google)	Anzahl indexierter Webseiten
2014	~30,000,000,000,000
2013	~22,000,000,000,000
2012	~17,000,000,000,000
2011	~12,000,000,000,000
2010	~6,000,000,000,000
2009	~3,000,000,000,000
2008	~1,000,000,000,000
2000	~1,000,000,000
1998	~26,000,000

- Statt viel Geld an Google für Werbung zu zahlen...
 - dafür sorgen, dass Webseite wichtig erscheint
 - **Linkfarmen, Spamdexing**
- Ältere Studien: 6-22% aller Seiten sind Spam (Spirin und Han, 2011)
- Aufgeschlüsselt nach Sprache:
 - 13,8% des englischen Webs sind Spam
 - 9% des japanischen
 - 22% des deutschen
 - 25% des französischen
- Aufgeschlüsselt nach Domaine:
 - 70% aller Webseiten in der *.biz-Domäne sind Spam
 - 20% in der *.com-Domäne sind Spam



[https://en.wikipedia.org/wiki/Spam_\(food\)#/media/File:Spam_can.png](https://en.wikipedia.org/wiki/Spam_(food)#/media/File:Spam_can.png)

- Websuchmaschinen sind teuer
 - Entwicklung, Hardware, Netzwerk, Strom
- Daher muss der Betreiber Geld verdienen, um den Betrieb langfristig aufrechtzuerhalten
- Mögliche Geschäftsmodelle
 - Werbefinanziert
 - Nutzerfinanziert
 - Mitgliederfinanziert
 - Datenfinanziert



Websuche

Geschäftsmodell: Werbefinanziert

- Geld für Werbeanzeigen auf Ergebnisseite
 - Populärstes Modell
 - Viele Nutzer notwendig und personalisierte Werbung

The screenshot shows a search engine results page for the query "investing money wisely". The search bar at the top indicates "About 1,040,000 results (0.29 seconds)". The results are divided into two main sections: "Organic Search Results" (highlighted with a green box and green arrows) and "Paid Search Results" (highlighted with a red box and red arrows).

Organic Search Results (Green Box):

- Investment Opportunities**
www.SDI.co.uk Scotland can deliver exactly what you are looking for. Let SDI help.
- Personal Online Investing**
www.Scottrade.com \$7 Online Equity Trades - No Limit. Invest With Confidence - Apply Now!
- Invest in Stocks for \$4**
www.ShareBuilder.com/Invest No Account or Investment Minimums. No Inactivity Fees. Start Today!
- How to Invest Small Amounts of Money Wisely - wikiHow**
Sep 29, 2010 ... How to Invest Small Amounts of Money Wisely ... Keep in mind that this is money for investing, and nothing else, so you need to realize that ...
www.wikihow.com/Invest-Small-Amounts-of-Money-Wisely - Cached - Similar
- Investing Money**
Investing Money wisely is something we'd all like to say that we do all the time. The problem is we sometimes slip up and wind up investing that money ...
www.money-zine.com/Investing/Investing-Money/ - Cached - Similar
- How to Become a Millionaire - The Top 5 Tips**
May 7, 2007 ... Now that you are saving money, you need to invest it wisely ... while others have done so by investing in a few stocks that exploded. ...
genxfinance.com/.../the-top-5-ways-to-become-a-millionaire/ - Cached - Similar
- 4 Ways To Invest Money Wisely - Financial Web**
In both difficult and prosperous economic times, it can be confusing to determine the best ways to invest money wisely. The key steps are to make a ...
www.finweb.com/investing/4-ways-to-invest-money-wisely.html - Cached - Similar
- Invest Wisely: Mutual Funds**
Jul 2, 2008 ... You can lose money investing in mutual funds. For more information about investing wisely and avoiding fraud, please check out the ...
www.sec.gov/investor/pubs/inwsmf.htm - Cached - Similar
- Basics - Personal Finance 101, Investing, Budgeting, Retirement ...**
See all Investing. Topics and Guides Knight Kiplinger. Follow these principles of smart money management to prosper in the long term. ...
www.kiplinger.com/basics/ - Cached - Similar

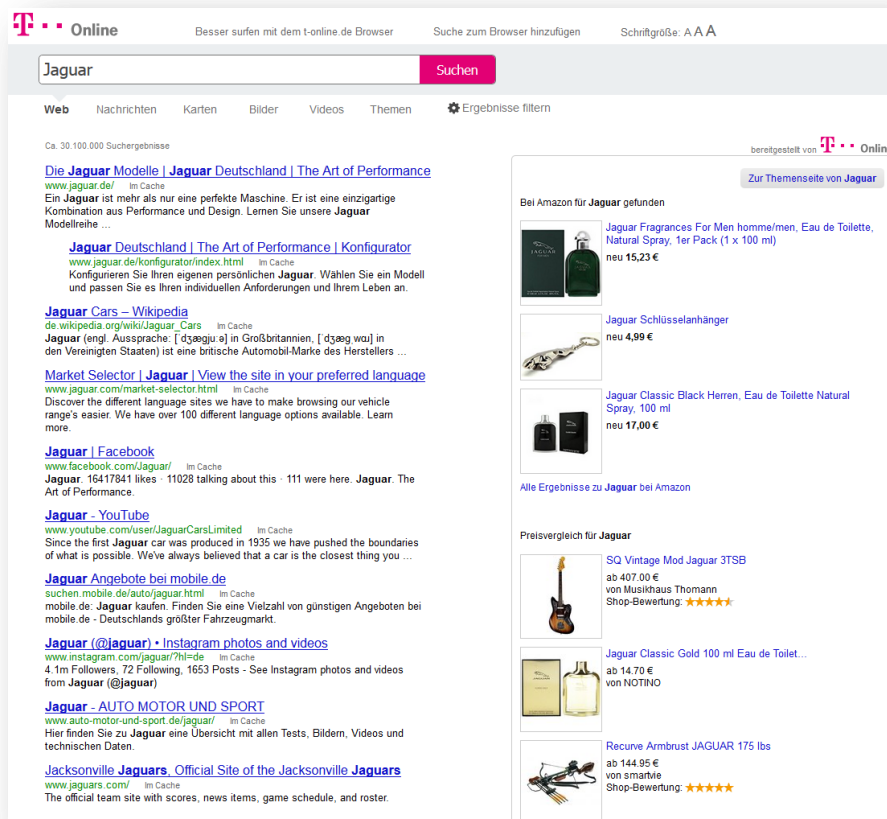
Paid Search Results (Red Box):

- Sponsored links**
- Merrill Lynch Wealth Mgmt**
Pay For Your Kids' Tuition & Still Invest in Retirement. Learn More.
www.TotalMerrill.com
- MarketSmith**
Institutional Grade Data - Powerful Stock Investment Tools. Learn More.
www.Marketsmith.com
- Dorfman is a Dork**
But His Trading Book is Great! Hedge Fund Trading Secrets Revealed
www.pristinepublishing.com
- Safe Money Strategies**
Get Market index returns Without Risk to your principal
www.safemoneystrategies.ws
- Investing 101**
Get the Basics-Be A Smart Investor Resources, Tools & Guides
www.SmartAboutMoney.org/Investing
- 1.31-1.61% Apply Online**
From AA+ GE Capital Corporation. Not An Offer Of Securities For Sale
www.geinterestplus.com
- 17.6% APY Returns \$5K Min**
\$5,000 Minimum Investment Serious Investors Only Please
www.InvestInCarNotes.com
- View Stock Behavior**
Free Online Trading Application. Research Investment Opportunities.
Kapital.com

Websuche

Geschäftsmodell: Nutzerfinanziert

- **Abomodell:** Nutzer bezahlen für Nutzung der Suchmaschine
 - Funktioniert nur, wenn die Suchmaschine wirklich unschlagbar gut ist
- Populärer: Suchmaschinentechnologie an andere Firmen vermieten
 - Z.B. t-online.de nutzt Googles Suchmaschine



- **Mitgliedermodell** finanziert sich über die Gemeinschaft
 - Nutzer helfen bei der Entwicklung
 - Spart Geld und erhöht Qualität
 - Betrieb wird durch Spenden und Werbeanzeigen finanziert
 - Z.B. Wikia Search (2009 eingestellt)



- Beim **Infomediary-Modell** ist die Nutzung der Suchmaschine kostenfrei, die anfallenden Nutzerdaten können aber zu Marketingzwecken ausgewertet werden
 - **Suchverhalten** der Nutzer wird analysiert
 - Erstellung von Benutzerprofilen
 - Vorhersagen von Suchtrends
 - **Verkauf der Daten** an andere Firmen
 - Optimierung der Werbestrategie
 - Legale Grauzone
 - Datenschutz, Privatsphäre
 - Z.B. ...



<http://vignette1.wikia.nocookie.net/villains/images/7/74/Dr-Evil-dr-evil-19031166-1280-1024.jpg>