Internetsuche und Google Page-Rank -

Wie wird was durch wen gefunden?

Magnus Knuth & Nadine Steinmetz

Hasso Plattner Institute for IT-Systems Engineering

University of Potsdam

HPI

Hasso Plattner Institute for IT Systems Engineering
Semantic Technologies & Multimedia Retrieval Research Group

- **Research Topics**

  - Semantic Web Technologies

  - Ontological Engineering

  - Information Retrieval

  - Multimedia Analysis & Retrieval

  - Social Networking

  - Data/Information Visualization

- **Research Projects:**

semex Semantic Media Explorer

THESEUS

yovisto Academic video search

yovistolabs

mediaglobe the digital archive

CONTENTUS Next Generation Multimedia Libraries

Themen

Suchindex

Suchmaschinen

Webseiten

Personalisierung

Suchergebnisranking

HTML

Suchmaschinenoptimierung

Multimedia

THE HISTORY OF SEARCH

4

▼1990

Archie
First search engine created by Alan Emtage, a student at McGill University in Montreal.

◄ Search Methods Before the Internet

▼1991

Birth of the Internet
Tim Berners-Lee launched his invention: the World Wide Web.

600 + Websites Online

10 Million + Internet Users

▼1993

Aliweb

Aliweb Launched
Allowed users to submit the locations of index files on their sites.

Wanderer
The first web crawler to measure the size of the web. Created by Matthew Gray

▼1994

WEBCRAWLER LIGHTNING FAST WEB SEARCH

LYCOS. Go get it!

WebCrawler
First search engine to index full web pages and allow searchers to search with any word.

Lycos
Search engine and web portal launched.

▼1995

YAHOO! Launched using a web directory, not a full text index.

THE HISTORY OF SEARCH

▼1998

4

Google Launched by Sergey Brin and Larry Page - PageRank was born.

DMOZ
Open Directory Project is launched and purchased by Netscape.

dmoz

Pay Per Click
Bill Gross developed the original PPC model.

SEO NOTE:
Most search engines moved to off-page criteria for their ranking algorithms.

MSN Search
Launched using Inktomi's search technology.

msn Search

▼1999

SES Conference
First Search Engine Strategies conference created by Danny Sullivan.

Search Engine STRATEGIES CONFERENCE & EXPO

279 Million + Internet Users

2.2 Million + Websites Online

▼2000

Yahoo!
Began using Google for search technology, but switched to its own technology in 2004.

Baidu
China's largest search engine launched.

Baidu 百度

THE HISTORY OF SEARCH

HPI Hasso Plattner Institut

▼ 2000

4

Yahoo!
Began using Google for search technology, but switched to its own technology in 2004.

Baidu
China's largest search engine launched.

PubCon
First event held in London.

Google
- Introduced the Google Toolbar.
- Matt Cutts joined the team.

AdWords Launched
AdWords is Google's advertising product and main source of revenue. Google's advertising revenue totaled $28 billion in 2010.

▼ 2001

The Exodus
Users abandoned other search engines such as HotBot, AltaVista and Excite in record numbers.

"Many SEOs have sleepless nights as we realize it is Google or bust." - Brett Tabke, Founder of WebmasterWorld

THE REIGN OF GOOGLE BEGINS

▼ 2002

Google
Announced the launch of Froogle, a product search engine.

Google Became a Verb:
The American Dialect Society chose it as the "most useful word of 2002." It was officially added to the Oxford English Dictionary on June 15, 2006.

Yahoo

THE HISTORY OF SEARCH

2002

Google
Announced the launch of Froogle, a product search engine.

Yahoo
Acquired Inktomi for $235 million and offered to buy Google for $3 billion.

Google Became a Verb:
The American Dialect Society chose it as the "most useful word of 2002." It was officially added to the Oxford English Dictionary on June 15, 2006.

2003

Blogger and WordPress
Blogging became popular and was utilized for SEO.

782 Million + Internet Users

38 Million + Websites Online

Google "Florida" Update
This update turned the SEO world upside down. The new algorithm was context based. After this update, it was emphasized that quality of content and HTML code were the most important factors for rankings on a search engine.

2004

seomoz.org

SEOmoz
Founded by Rand Fishkin, first a blog, SEOmoz grew to become one of the largest providers of SEO Tools.

Google Local
Google Local introduced to offer relevant neighborhood business listings, maps and directions.

2005

THE HISTORY OF SEARCH

HPI Hasso Plattner Institut

▼ 2009

bing Windows Live Search was rebranded and became Bing.

Google
- Launched Google Caffeine.
- Incorporated Twitter data into search results. ▼
- Introduced Rich Snippets.
- Updated their brand.

twitter

New Canonical Tag
Supported by all three major search engines.

▼ 2010

2 Billion + Internet Users

240 Million + Websites Online

Google
Launched Instant Previews.

facebook

Facebook
Launched Open Graph API, which included the new "Like" button.

Yahoo! and Microsoft
Yahoo partnered with Microsoft to use Bing search technology.

▼ 2011

Facebook
Received patent on "curated search," which involved ranking search results according to a link's popularity with a user's social graph.

Google Farmer/Panda Update
Google's algorithm update devalued content farms like ezinearticles.com and others.

THE HISTORY OF SEARCH

Facebook

Received patent on "curated search," which involved ranking search results according to a link's popularity with a user's social graph.

Google +1

Google kicked social search into high gear with its +1 button, similar to Facebook's "Like" button. In 2011, 25 percent of bonuses may be based on the success or failure of Google's social efforts.

Google Farmer/Panda Update

Google's algorithm update devalued content farms like ezinearticles.com and others.
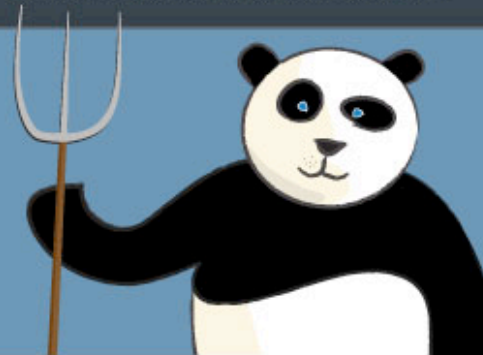
Search Engine Market Share

Legend: Google, Yahoo, MSN, Bing
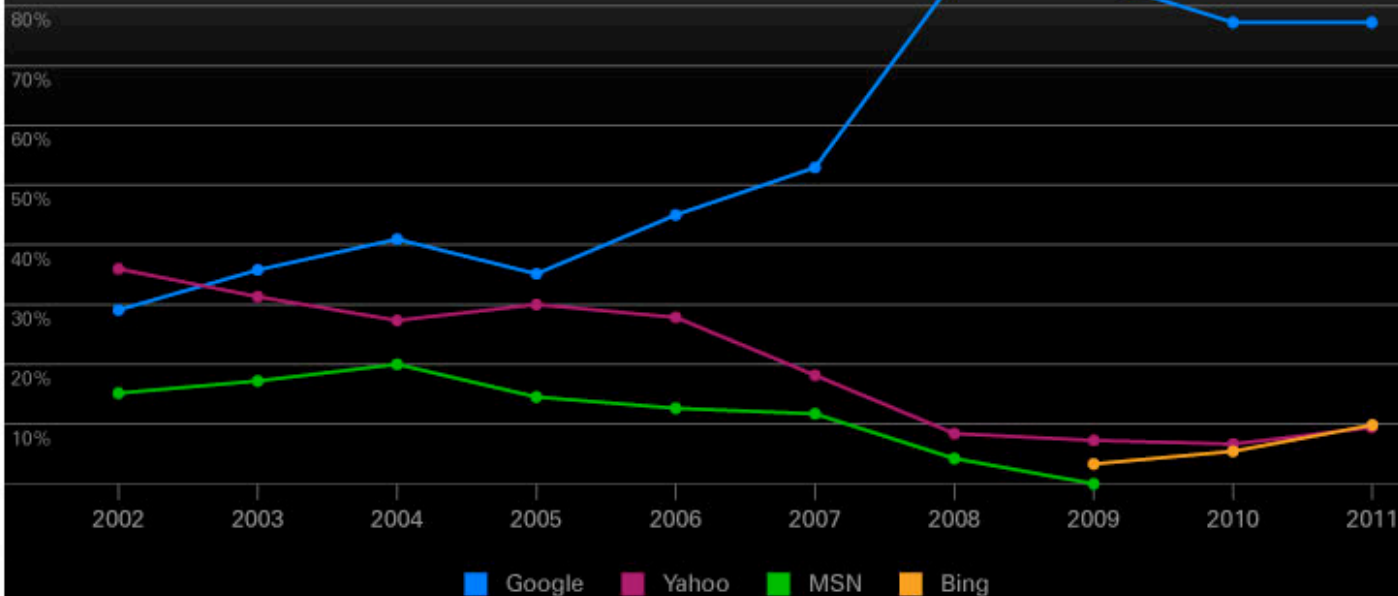
Was ist ein Index?

Suchindex - Arten

Forward-Index: Zuordnung von Termen zu Webseiten (Dokument muss nicht komplett durchsucht werden, Terme sortiert)

Dok1	Die, Katze, macht, miau
Dok2	Die, Kuh, macht, muh
Dok3	Die, Erbse, grün, ist
Dok4	Die, Kuh, lacht

Invertierter Index: Zuordnung von Webseiten zu Termen (findet schnell Dokumente zu Suchterm)

Die	Dok1, Dok2, Dok3, Dok4
Kuh	Dok2, Dok4
macht	Dok1, Dok2
muh	Dok2

Dokument-Term-Matrix: zweidimensionale Matrix (Anzahl des Auftreten in Dokument)

	Dok1	Dok2	Dok3	Dok4
Die	1	1	1	2
Kuh		1		1
macht	1	1		
muh		1		

1 Google sends out Web crawlers, known as Googlebots, that go through all the Web pages on the World Wide Web.

GOOGLEBOTS

2 These bots are made up of millions of computers that surf the Web much like how people do it, except they do it much more quickly. To make sure that they have the most current sites, these Googlebots are working all the time, crawling and re-crawling websites.
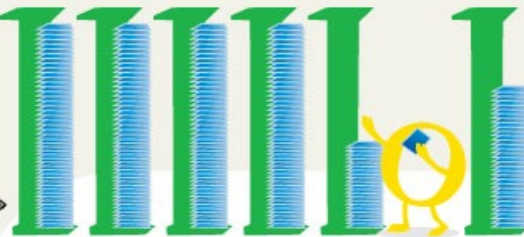
24

3 How often a site is crawled depends on how quickly its content changes. For example, a newspaper site that has frequent updates will get crawled more often than a static company site. Webmasters can also indicate how often they want their sites to be crawled.
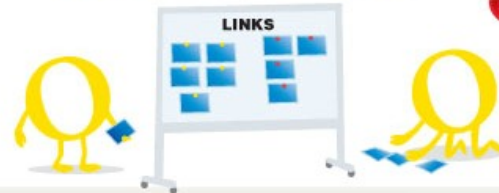
D-F

INDEX DATABASE

5 Copies of all these pages are stored in Google's index database. The index works like the contents page of a book and contains all the possible search terms.

LINKS

4 The bots also detect all links on the site, which are then put into a queue system for crawling later on. These links are also important in deciding the search ranking of a site.

Query

Query

7 At the same time, a short description of the pages is generated for the search results page.

SEARCH RESULTS

6 When a user types in a query in the search engine, the index will match the query with all the webpages in which the query terms appear and then grab the pages for the search results.

COLD WARM HOT!

8 Web pages are ranked in terms of relevance before they are displayed on the results page. Google considers some 200 factors when ranking sites. One of them is PageRank, which takes into consideration how many sites are linked to a webpage and the quality of the linking sites.

Tham Yuen-C, Quek Hong Shin: „How Google works"

Indexierung

- Probleme bei der Indexierung

  - Wortgrenzen (deutsch, englisch wenig problematisch): Wo hört ein Wort auf?
    Zentraleuropa oder **Zentrale Uropa**?

  - Sprachabhängigkeit: In welcher Sprache liegt das Dokument vor?
    Boot (de) oder **boot (en)**?

  - diverse Datenformate: Welche Zeichenkodierung verwendet das Dokument?
    ü = ü = ¨u = \"u

Welche Informationen einer Webseite landen im Index?

Welche Informationen einer Webseite landen im Index?

http://www.naturkundemuseum-berlin.de/en/ausstellungen/saurierwelt/kentrosaurus/

http://www.dinosinart.com/

- Ansicht Webbrowser / HTML-Code

- Welche Information ist relevant?

- Welche Information ist wichtiger als andere?
  - Section Recognition

- Welche Information kann nicht indiziert werden?

Schlagwortbasierte Suche

Suchaufgaben

- **Kultur**
  Die Nationalflaggen aller Länder der Erde haben eine gemeinsame geometrische Form, außer eine. *Welches Land ist das?*

- **Geographie**
  Forscherteams aus Kushiro, Pevek und Namsos erreichen den Nordpol und überprüfen ihre Uhren. Wenn sie entsprechend 11:42 Uhr, 12:17 Uhr und 3:00 Uhr ankamen, *wer war als Erstes da?*

- **Astronomie**
  Diese Strahlung ist gemeinhin bekannt zum Erwärmen von Speisen, aber lieferte ebenfalls den Beleg für eine Theorie von Georges Lemaître. *Von welchem astronomische Ereignis zeugt diese Strahlung?*

13

- Findet ein Wort oder eine Wortkombination für die Google kein Ergebnis liefert!

Venezianischer Zwergspringer

- Baue diese Wortkombination in eine Webseite ein.

  □ E-Mail mit URL an magnus.knuth@hpi.uni-potsdam.de

- Welche Webseite wird am Ende des Schülerkollegs am höchsten gerankt?

semex
Semantic Media Explorer

Kontakt:

Nadine Steinmetz & Magnus Knuth

Hasso-Plattner-Institut für Softwaresystemtechnik

Universität Potsdam

Prof.-Dr.-Helmert-Str. 2-3

D-14482 Potsdam

E-Mail: vorname.nachname@hpi.uni-potsdam.de

Danke für eure Aufmerksamkeit.

Suchaufgaben II

- **Kultur**

  Rembrandt malte ein Bild von einem Philosophen, der eine Büste eines griechischen Poeten anschaut. Ein Gold-Medaillon auf der Büste zeigt einen anderen berühmten Griechen. *Wen?*

- **Geschichte**

  Im April 1896 habe ich einen Mann erschossen. Acht Monate davor hatte dieser einen anderen Mann erschossen, der wiederum 17 Jahre davor 42 Männer erschossen haben soll. *Wen habe ich erschossen?*

- **Geographie**

  Die zwei Länder, auf deren Territorium früher die Hamangia lebten, sind durch einen Grenzfluss voneinander getrennt. *Wie lang ist dieser Fluss insgesamt?*