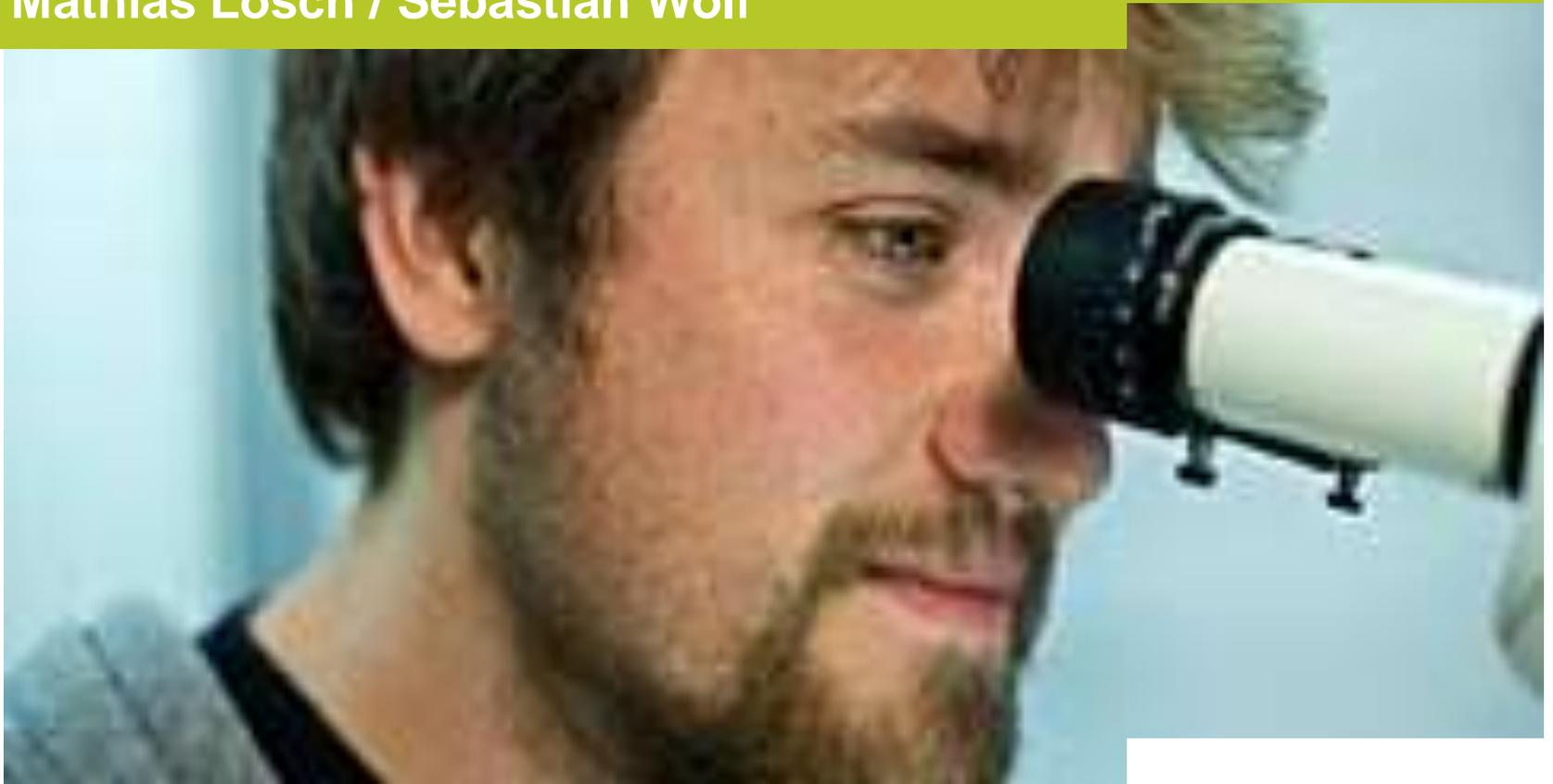


# Grundlagen der Suchtechnologie

Kolloquium Wissensinfrastruktur, 28.05.2010

Mathias Lösch / Sebastian Wolf



# Inhalt

- [Historie](#)
- [Indexierung](#)
- [Suchtechnologien](#)
- [Ranking-Verfahren](#)
- [Suchmaschinenoptimierung](#)
- [Ausblick](#)

# Historie

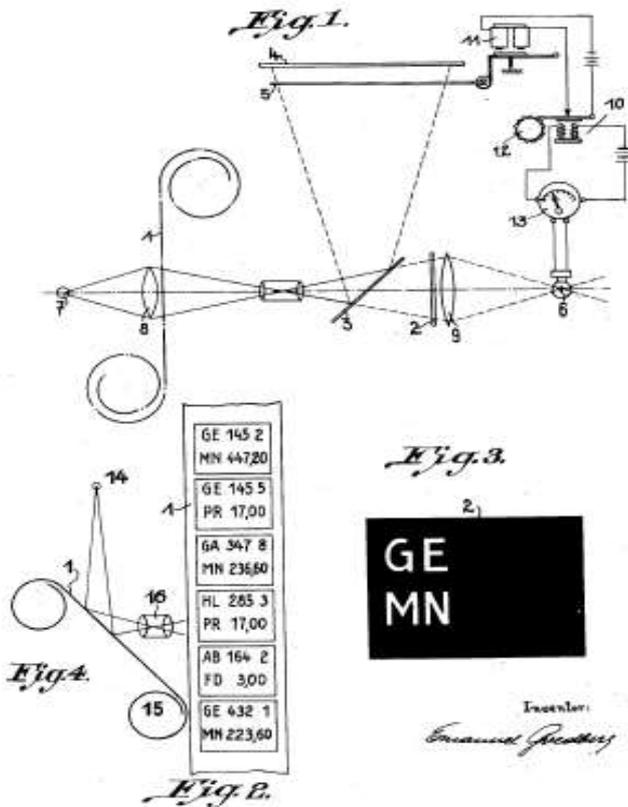


## Register / Index (1614)

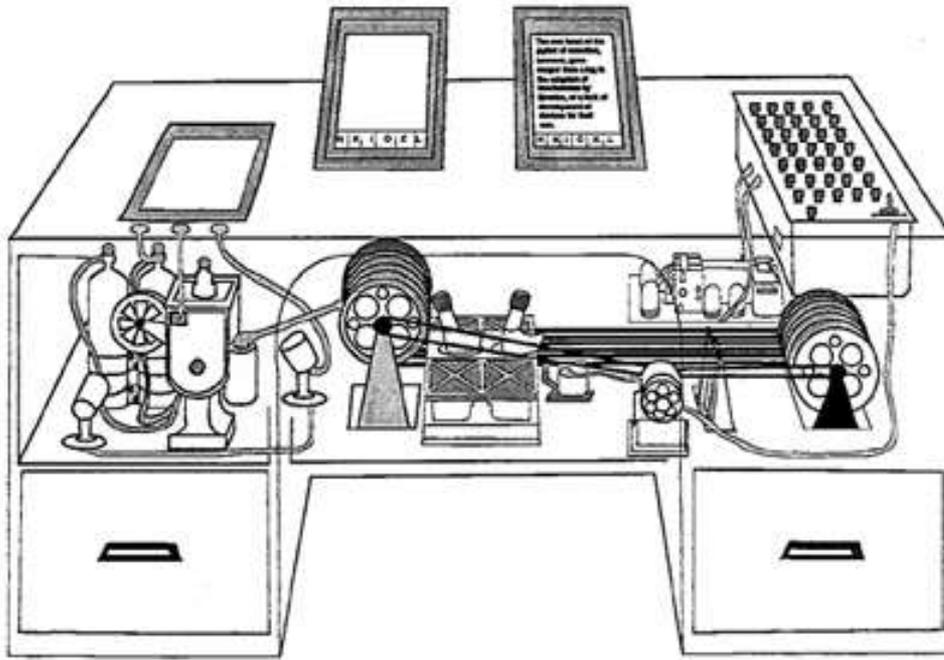
Antonio Zara (1574-1621), Bischof von Petina, fügte seiner Enzyklopädie *Anatomia ingeniorum et scientiarum* erstmals einen umfangreichen Index an

# Historie

**Statistische Maschine (1931)**  
 Sollte Metadaten auf Rollen von Mikrofilm mit Hilfe von Fotozellen und Mustererkennung durchsuchen können. Vorgestellt von Emanuel Goldberg 1931 auf dem „VIII. Internationalen Kongreß für wissenschaftliche und angewandte Photographie“



# Historie



Quelle: <https://atlas.colorado.edu/~hofmocke/digitalpoetry/memex.html>

## Memex (1945)

Als Wissensfindungs- und Verwertungssystem konzipierter Kompakt-Analog-Rechner, der 1945 von Vannevar Bush im Artikel „As We May Think“ fiktiv vorgestellt wurde

# Historie

## The Answer Machine

You're doing your homework.  
You're stuck and you need some answers.  
So you get help from your answer machine.  
On a table next to you is part of the machine—  
a typewriter keyboard.  
When you push the correct keys,  
your questions will be answered  
on a screen on the wall.



You might ask (using the keyboard):

WHO INVENTED THE PHONOGRAPH?

And the machine would answer:

THOMAS A.  
EDISON

## The Answer Machine (1964)

Theoretischer Entwurf einer Suchmaschine  
(in: Childcraft Vol. 6  
How Things Change)

# Historie

## DIALOG (1972)

Der Datenbankanbieter Dialog stellte 1972 mit dem DIALOG Information Retrieval Service das weltweit erste kommerzielle Angebot mit Zugriff auf eine Online-Datenbank vor.

```
UTILITY
OBERFLAENCHENWASSER OR STILLGWASSER OR SEE#1 OR TEICH#1 OR FEUCHTGEBIETE#
RENATURIERUNG OR BIOTOPGESTALTUNG OR BIOTOPENSCHUTZ#
1 AND 2#

+SIGN-ON 11.09.52 26.09.06
D-S/UTILITY/1976 - 10. SEPTEMBER 1986 SESSION 27
COPYRIGHT BY UMWELTBUNDESAMT D-1000 BERLIN 33

D-S - SEARCH MODE - ENTER QUERY
1_1 OBERFLAENCHENWASSER OR STILLGWASSER OR SEE#1 OR TEICH#1 OR FEUCHTE
#ITE

RESULT 2568

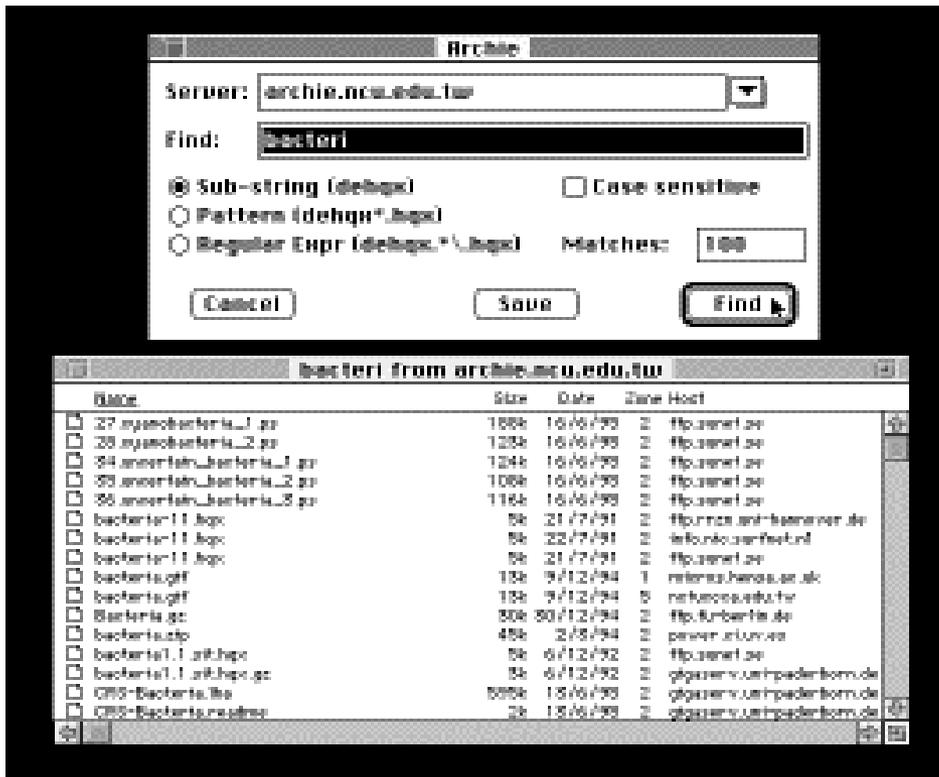
2_1 RENATURIERUNG OR BIOTOPGESTALTUNG OR BIOTOPENSCHUTZ

RESULT 687

2_1 1 AND 2

Blattstopp mit 1 und 2 Abschließen mit +^ Zurück: Jede andere Eingabe.
(66) MITSPICHERN (PRINT) MITDRUCKEN CHAB LOCK 10.09.73
```

# Historie



**Archie (1990)**  
Archie war eine Suchmaschine, die speziell für das Indizieren von FTP-Archiven entwickelt wurde. Gesucht werden konnte nach (Teilen von) Dateinamen.

# Historie

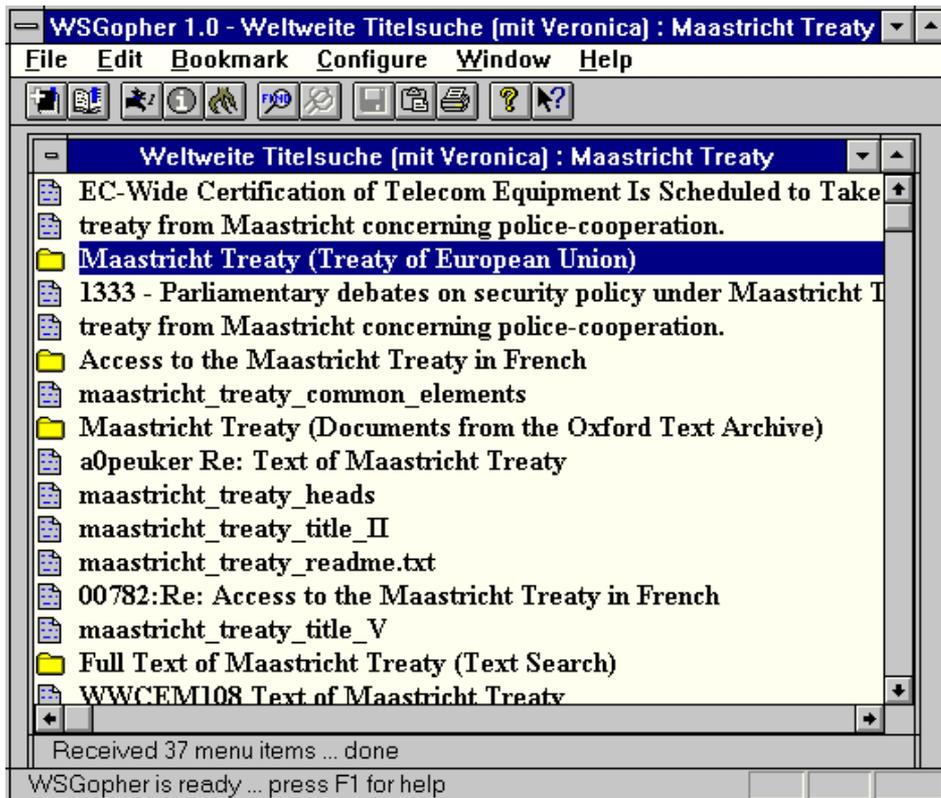


## WAIS (1991)

„Wide Area Information Server System“.

Innerhalb des WAIS-Systems wurden Dateien gelistet, die im Internet erreichbar sind. WAIS ermöglicht erstmals eine Volltextsuche innerhalb des Datenbestandes.

# Historie



**Veronica (1992)**  
„Very Easy Rodent-Oriented Netwide Index to Computerized Archives“. Suchdienst für „Gopher“-Webseiten (Titel bzw. Dateinamen).

# Historie



The screenshot shows the WebCrawler search engine interface. At the top, there is a logo for WebCrawler with the tagline "Search before you surf!". Below the logo are several icons: a magnifying glass for "Search", a folder for "Browse", a star for "Special", a plus sign for "Add URL", and a question mark for "Help".

The main search area includes a text input field, a "Search" button, and a dropdown menu for "titles" and a "25" results limit. Below the search field, there is an example search query: "Example: 'Alien Abduction' UFO Roswell [Search tips](#)".

Below the search area, there is a "WebCrawler SELECT" section with a search result: "Nobody covers sports like SportsLine." Below this, there is a list of categories: "Choose one of these categories: Arts & Literature - Business - Chat - Computers - Daily News - Education - Entertainment - Government - Health & Medicine - Internet - Kids & Families - Life & Culture - Personal Finance - Recreation - Reference Desk - Science - Sports - Travel".

At the bottom, there is a link to "Gossip" section: "Get the latest dirt in our new [Gossip](#) section!".

At the very bottom, there is a navigation bar with links: "Search · [Browse](#) · [Special](#) · [Add URL](#) · [Help](#)".

Copyright © 1996 America Online, Inc. [Disclaimer](#)

**Webcrawler (1994)**  
Erste Volltext-Internet-Suchmaschine, die für die Ergebnisanzzeige eine Relevanzbewertung (Ranking) vornahm

# Historie



Search the web using Google!

10 results



Google Search

I'm feeling lucky

*Index contains ~25 million pages (soon to be much bigger)*

## About Google!

[Stanford Search](#) [Linux Search](#)

Get Google! updates monthly!

your e-mail

Subscribe

[Archive](#)

Copyright ©1997-8 Stanford University

**Google (1998)**  
Führte eine neue  
Form des Rankings  
(Link-Popularität) ein

# Indexierung

Dieser Teil ist als separate Präsentation abrufbar:

[Starten](#)

# Suchtechnologien

- Abgleich von Zeichenketten
- Reguläre Ausdrücke
- „Fuzzy“-Suche
- Phonetische Suche
- Semantische Suche

# Suchtechnologien: Abgleich von Zeichenketten

- Das gesuchte Wort wird mit dem im Index verglichen
- Über die Phrasensuche können Wortgruppen abgeglichen werden („Universität Bielefeld“)
- Bei der Eingabe mehrerer Wörter, wird i.d.R. mit „UND“ verknüpft (teilweise mit „ODER“)
- Weitere Verknüpfungsmöglichkeit z.B. durch Verwendung Boolescher Operatoren (AND, OR, NOT)

# Suchtechnologien: Reguläre Ausdrücke

- Ersetzen eines oder beliebig vieler Zeichen bzw. eines Zeichenbereichs (a-z, 0-9, f-q, B oder D, Leerzeichen, Tabulatoren oder anderer Steuerzeichen)
- Für große Datenmengen ungeeignet (Suchzeiten)
- Suchmaschinen beherrschen (wenn überhaupt) nur das „Wildcard“-Zeichen am Wortende (Trunkierung; ersetzt beliebig viele Zeichen)

# Suchtechnologien: Fuzzy-Suche

- Normierung von Umlauten / Sonderzeichen  
(ä = ae, á = a)
- Rechtschreibkontrolle anhand eines Wörterbuchs  
(Eifelturm = Eiffelturm)
- Ermittlung von Synonymen anhand eines Wörterbuchs  
(Fernsehgerät = Fernseher)
- Unscharfe-UND-Verknüpfung („Fuzzy-AND-Search“)

# Suchtechnologien: Fuzzy-Suche

- Stemming / Lemmatisierung = verschiedene morphologische Varianten eines Wortes auf ihren gemeinsamen Wortstamm zurückgeführt
- Numerus (Bibliotheken = Bibliothek)
- Flexion (Houses = Haus / schneller = schnell / geworden = werden)
- Komposita (Fußballstadion = Fußball + Stadion)

# Suchtechnologien: Fuzzy-Suche

- „Unschärfer“ Abgleich des Suchbegriffs (Levenshtein-Distanz / N-Gramm-Analyse)
- Levenshtein-Distanz = minimale Anzahl von Einfüge-, Lösch- und Ersetz-Operationen um die erste Zeichenkette in die zweite umzuwandeln
- Beispiel: Die Begriffe Tier und Tor haben eine Levenshtein-Distanz von 2:  
Tier → Toer (Ersetze i durch o) → Tor (Lösche e)

# Suchtechnologien: Fuzzy-Suche

- N-Gramm-Analyse = Berechnung der Wahrscheinlichkeit, dass auf eine bestimmte Buchstaben- oder Wortreihenfolge ein bestimmter Buchstabe oder ein bestimmtes Wort folgen wird
- Durch Abgleich mit einem Referenzdokument lassen sich Cluster bilden
- Je näher ein Dokument am Referenzdokument liegt, umso wahrscheinlicher ist, dass sich der Inhalt um dessen Thema dreht

# Suchtechnologien: Phonetische Suche

- Gleichklingende Wörter werden mit einer identischen Zeichenfolge kodiert
- Bei einer Suche nach einem Wort, wird auch das ähnlich klingende Wort gefunden (oder als Suchbegriff vorgeschlagen)

# Suchtechnologien:

## Phonetische Suche (Beispiel „Soundex“)

- Ähnliche Laute besitzen den gleichen Code (im englischen werden z.B. B, F, P und V mit der Ziffer „1“ codiert)
- Vokale, Umlaute und bestimmte Konsonanten werden ignoriert
- Der Anfangsbuchstabe des Wortes bleibt erhalten
- Neben dem Anfangsbuchstaben werden maximal 3 Konsonanten kodiert, ggf. wird mit 0 aufgefüllt

# Suchtechnologien:

## Phonetische Suche (Beispiel „Soundex“)

- Ergebnisse sind oft zielführend, aber manchmal auch irreführend oder unsinnig
- **Google** = **G240**
- **Googlebombe** = **G241**
- **Gugelhupf** = **G241**
- Für jede Sprache braucht man einen eigenen Code (Deutsch = „Kölner Phonetik“)

# Suchtechnologien: Semantische Suche

- Klassifizierung von Dokumenten mit Hilfe texttechnologischer Verfahren
- Ermittlung des Kontextes, in dem ein Suchbegriff steht (z.B. Laster = LKW / Schlechte Angewohnheit)
- Ermittlung der Intention des Nutzer (wonach sucht der Nutzer)
- Problem: Bei kurzen Suchanfragen nicht geeignet

# Ranking-Verfahren

- Formale Sortierung
- On-Page-Faktoren
- On-Site-Faktoren
- Link-Faktoren
- Eigenschaften und Verhalten der Nutzer

# Ranking-Verfahren: Formale Sortierung

- Alphabetische Sortierung nach Titel oder Autor
- Chronologische Sortierung nach Erscheinungsdatum
- Nur in Bereichen sinnvoll und möglich, die über entsprechende Metadaten verfügen

# Ranking-Verfahren: On-Page-Faktoren

- Häufigkeit und Position (Dichte, Abstand) der Terme
  - Funktion (URL, HTML-Auszeichnung: Titel, Überschrift, Linktext)
  - Format (Schriftgröße und –farbe)
- Hohes Missbrauchspotential

# Ranking-Verfahren: On-Site-Faktoren

- Analyse globaler Faktoren der jeweiligen Domain
- Alter der Domain  
Art der Domain (Kommerziell / Wissenschaftlich)  
Thematische Ausrichtung  
Gesamtzahl der indexierten Seiten  
Linkpopularität der gesamten Domain

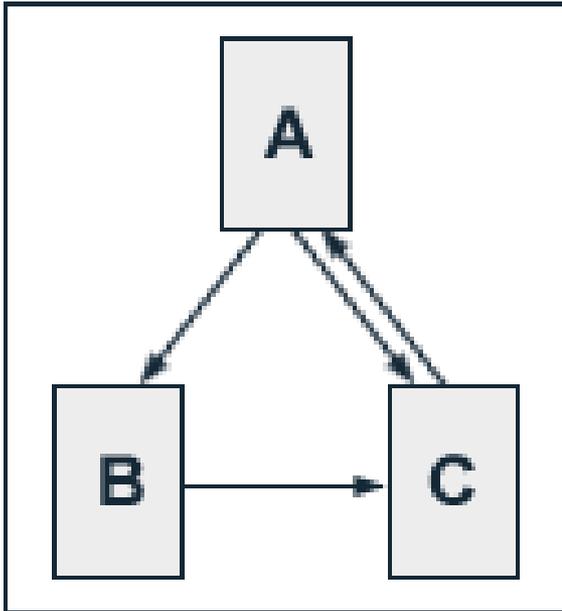
# Ranking-Verfahren: Link-Faktoren

- Ranking basierend auf der Analyse der Referenzstruktur im Web
- Annahme: Link stellt ein Qualitätsurteil dar (ähnlich einem Zitat / Zitationsanalyse)
- PageRank: Ermittlung der Wichtigkeit einzelner Dokumente durch Analyse der Verweisstruktur aller indexierten Webseiten

# Ranking-Verfahren: PageRank (einfache Fassung)

- $PR(A) = (1-d) + d (PR(T1)/C(T1) + \dots + PR(Tn)/C(Tn))$
- $PR(A)$  = PageRank der Seite A
- $d$  = Dämpfungsfaktor, wobei  $0 \leq d \leq 1$  ist  
(i.d.R. 0,85)
- $PR(Tn)$  = PageRank der Seite  $Tn$ , von denen ein Link auf die Seite A zeigt
- $C(Tn)$  = Gesamtanzahl der Links von Seite  $Tn$

# Ranking-Verfahren: PageRank (einfache Fassung, $d = 0,5$ )



- $PR(A) = 0.5 + 0.5 PR(C)$   
 $PR(B) = 0.5 + 0.5 (PR(A) / 2)$   
 $PR(C) = 0.5 + 0.5 (PR(A) / 2 + PR(B))$
- Bei Ausgangswert 1 für alle Seiten ergibt sich nach 13 Iterationen
- $PR(A) = 14/13 = 1.07692308$   
 $PR(B) = 10/13 = 0.76923077$   
 $PR(C) = 15/13 = 1.15384615$

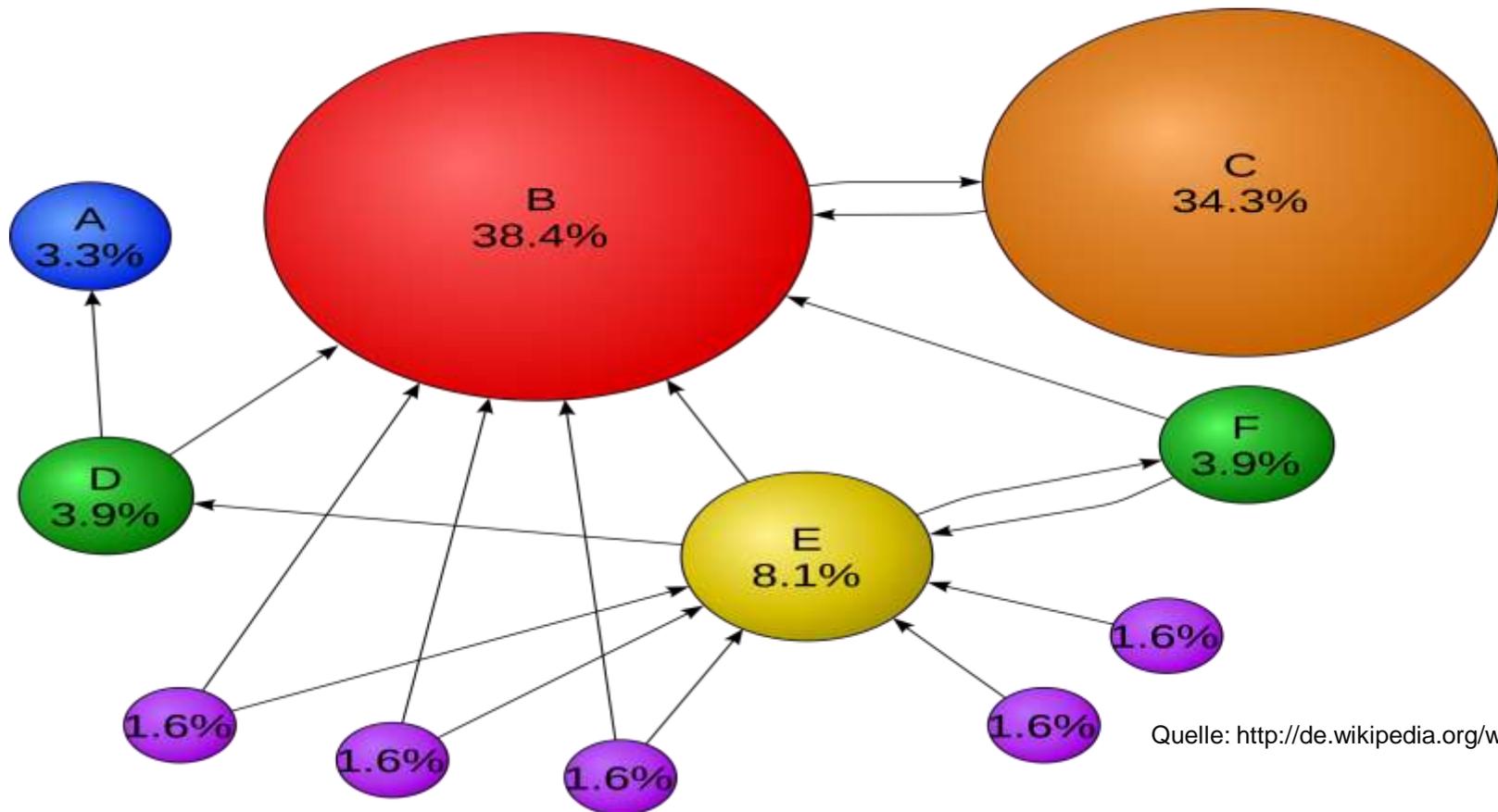
Quelle: <http://pr.efactory.de/d-pagerank-algorithmus.shtml>

# Ranking-Verfahren: PageRank-Wert und echter PageRank

Toolbar-PR	Tatsächlicher PR		
0/10	0.15	-	0.9
1/10	0.9	-	5.4
2/10	5.4	-	32.4
3/10	32.4	-	194.4
4/10	194.4	-	1,166.4
5/10	1,166.4	-	6,998.4
6/10	6,998.4	-	41,990.4
7/10	41,990.4	-	251,942.4
8/10	251,942.4	-	1,511,654.4
9/10	1,511,654.4	-	9,069,926.4
10/10	9,069,926.4	-	$0.85 \times N + 0.15$



# Ranking-Verfahren: Besuchs-Wahrscheinlichkeit



Quelle: <http://de.wikipedia.org/wiki/PageRank>

# Ranking-Verfahren: Link-Faktoren

- Linktexte von eingehenden Links werden ebenfalls indexiert und in das Ranking einer Webseite einbezogen
- Eine Webseite ist auch dann auffindbar bzw. wird als Treffer angezeigt, wenn das gesuchte Wort lediglich im Linktext eines externen Verweises, nicht aber auf der Seite selbst vorkommt
- Missbrauchspotential: „Google Bombing“

# Ranking-Verfahren: „Google Bombing“



blödzeitung

About 16,900 results (0.31 seconds)

 Everything

 Images

 More

All results

Sites with images

 More search tools

Tip: [Search for English results only](#). You can specify your search language in [Preferences](#)

[Aktuelle Nachrichten - Bild.de](#) ☆ - [ [Translate this page](#) ]

BILD.de: Die Seite 1 für aktuelle Nachrichten, Bilder und Videos aus den Bereichen News, Wirtschaft, Politik, Show, Sport, und Promis.

[www.bild.de/](#) - 8 minutes ago - [Similar](#)

<a href="#">Sport</a>	<a href="#">Erotik</a>
<a href="#">News</a>	<a href="#">Bundesliga</a>
<a href="#">Unterhaltung</a>	<a href="#">Schlagzeilen des Tages</a>
<a href="#">Fußball</a>	<a href="#">BILD-Girl</a>

[More results from bild.de »](#)

# Ranking-Verfahren:

## Eigenschaften und Verhalten der Nutzer

- Geografische Zuordnung des Users anhand der IP-Adresse (Land, Region, Stadt)
- Darstellung und Reihenfolge in Abhängigkeit der geografischen Zuordnung
- Suchergebnis wird mit Karten angereichert (Standorte von Firmen, Kneipen, Bibliotheken)
- Bei persönlichem Login zusätzlich Analyse der Suchhistorie des Anwenders

# Ranking-Verfahren: Eigenschaften und Verhalten der Nutzer



Google   [Erweiterte Suche](#)

Suche:  Das Web  Seiten auf Deutsch  Seiten aus Deutschland

Web [+ Optionen anzeigen...](#) Ergebnisse 1 - 10 von ungefähr 1.210.000

[UB Heidelberg - Universitätsbibliothek Heidelberg](#) ☆  
Vorstellung der Bibliothek, ihrer Geschichte, Sammelgebiete und Ausstellungen. Mit Informationen zu den Dienstleistungen und Online-Katalogen.  
[Heidi](#) - [E-Journals](#) - [Elektronische Medien](#) - [Kontakt & Öffnungszeiten](#)  
[www.ub.uni-heidelberg.de/](#) - [Im Cache](#) - [Ähnlich](#)

[Universitätsbibliothek](#) ☆  
Homepage der **Universitätsbibliothek** München. Hier finden Sie Bücher, E-Books, Datenbanken...Medien aller Art.  
[Opac](#) - [E-Medien](#) - [Bibliotheken](#) - [Kontakt](#)  
[www.ub.uni-muenchen.de/](#) - [Im Cache](#) - [Ähnlich](#)

[Universitätsbibliothek Bielefeld](#) ☆ - 2 Besuche - 19. Mai  
Die UB bietet über 2 Mio. Bücher und Zeitschriften, 95% davon frei zugänglich und ein ständig wachsenden Angebot wissenschaftlich relevanter ...  
[www.ub.uni-bielefeld.de/](#) - [Im Cache](#) - [Ähnlich](#)

Als eingeloggter Nutzer erscheint die UB Bielefeld auf Platz 3

# Ranking-Verfahren: Eigenschaften und Verhalten der Nutzer



The screenshot shows a Google search interface. The search bar contains the text 'universitätsbibliothek'. Below the search bar, there are radio buttons for 'Suche:  Das Web  Seiten auf Deutsch  Seiten aus Deutschland'. To the right of the search bar is a 'Suche' button and a link for 'Erweiterte Suche'. Below the search bar, there is a header for the search results: 'Web [+ Optionen anzeigen...](#) Ergebnisse 1 - 10 von ungefähr 1.210.001'. The first three search results are listed below:

[UB Heidelberg - Universitätsbibliothek Heidelberg](#)  
Vorstellung der Bibliothek, ihrer Geschichte, Sammelgebiete und Ausstellungen. Mit Informationen zu den Dienstleistungen und Online-Katalogen.  
[Heidi](#) - [E-Journals](#) - [Elektronische Medien](#) - [Kontakt & Öffnungszeiten](#)  
[www.ub.uni-heidelberg.de/](http://www.ub.uni-heidelberg.de/) - [Im Cache](#) - [Ähnlich](#)

[Universitätsbibliothek](#)  
Homepage der **Universitätsbibliothek** München. Hier finden Sie Bücher, E-Books, Datenbanken...Medien aller Art.  
[Opac](#) - [E-Medien](#) - [Bibliotheken](#) - [Kontakt](#)  
[www.ub.uni-muenchen.de/](http://www.ub.uni-muenchen.de/) - [Im Cache](#) - [Ähnlich](#)

[Universitätsbibliothek Leipzig - Startseite](#)  
Zugang zu Online-Katalogen und Datenbanken. Angabe von Öffnungszeiten, Serviceangeboten und Ansprechpartnern.  
[www.ub.uni-leipzig.de/](http://www.ub.uni-leipzig.de/) - [Im Cache](#) - [Ähnlich](#)

An arrow points from the text 'Als anonym Nutzer erscheint ein anderer Treffer auf Platz 3' to the third search result, 'Universitätsbibliothek Leipzig - Startseite'.

Als anonym Nutzer  
erscheint ein anderer  
Treffer auf Platz 3

# Ranking-Verfahren: Eigenschaften und Verhalten der Nutzer

Google   [Erweiterte Suche](#)

Suche:  Das Web  Seiten auf Deutsch  Seiten aus Deutschland

Web [+ Optionen anzeigen...](#) Ergebnisse 1 - 10 von ungefähr 12

**Bibliothek – Wikipedia**

Eine **Bibliothek** (griechisch βιβλιοθήκη „Büchersammlung“) oder Bücherei ist eine Dienstleistungseinrichtung, in deren Zentrum die publizierte Information in ...  
[de.wikipedia.org/wiki/Bibliothek](http://de.wikipedia.org/wiki/Bibliothek) - [Im Cache](#) - [Ähnlich](#)

**Deutsche Bibliotheken Online - hbz — Willkommen beim hbz**

Alphabetische Liste mit allen deutschen **Bibliotheken**, die im Netz vertreten sind.  
[www.hbz-nrw.de/produkte\\_dienst/germlst/index.html](http://www.hbz-nrw.de/produkte_dienst/germlst/index.html) - [Ähnlich](#)

**DNB, Deutsche Nationalbibliothek - Home**

Im Lesesaal der Anne-Frank-Shoah-**Bibliothek** der Deutschen Nationalbibliothek in Leipzig ist seit kurzem eine Bronzebüste von Anne Frank zu sehen, ...  
[www.d-nb.de/](http://www.d-nb.de/) - [Im Cache](#) - [Ähnlich](#)

**Lokale Branchenergebnisse für bibliothek im Umkreis von Bielefeld** - [Ort ändern](#)



- A** [Stadtbibliothek Bielefeld](#)  
[www.bibliotheken-in-owl.de](http://www.bibliotheken-in-owl.de) - 0521 5150-00 - [6 Bewertungen](#)
- B** [Bibliothek Verl](#)  
[www.bibliothek.verl.de](http://www.bibliothek.verl.de) - 05246 92523-0 - [Mehr](#)
- C** [Universitätsbibliothek Bielefeld](#)  
[www.ub.uni-bielefeld.de](http://www.ub.uni-bielefeld.de) - 0521 106-4051 - [Mehr](#)
- D** [Leopoldshöhe](#)  
[www.leopoldshoeh.de](http://www.leopoldshoeh.de) - 05208 991-330 - [Mehr](#)

Auch als anonymen Nutzer  
 Erkennt Google meinen  
 Standort anhand der IP  
 Und gibt passende  
 Ergebnisse aus

# Suchmaschinen-Optimierung (SEO)

- Optimierung einer Webseite hinsichtlich der Indexierung und der Relevanzbewertung einer Suchmaschine
- Aussagekräftige Titel, Überschriften und Linktexte
- Optimierung der internen und externen Verlinkungsstruktur
- ASEO = Academic Search Engine Optimization (Optimierung von PDFs für Google Scholar)

# Suchmaschinen-Optimierung (SEO)

URL: [www.ub.uni-bielefeld.de](http://www.ub.uni-bielefeld.de)

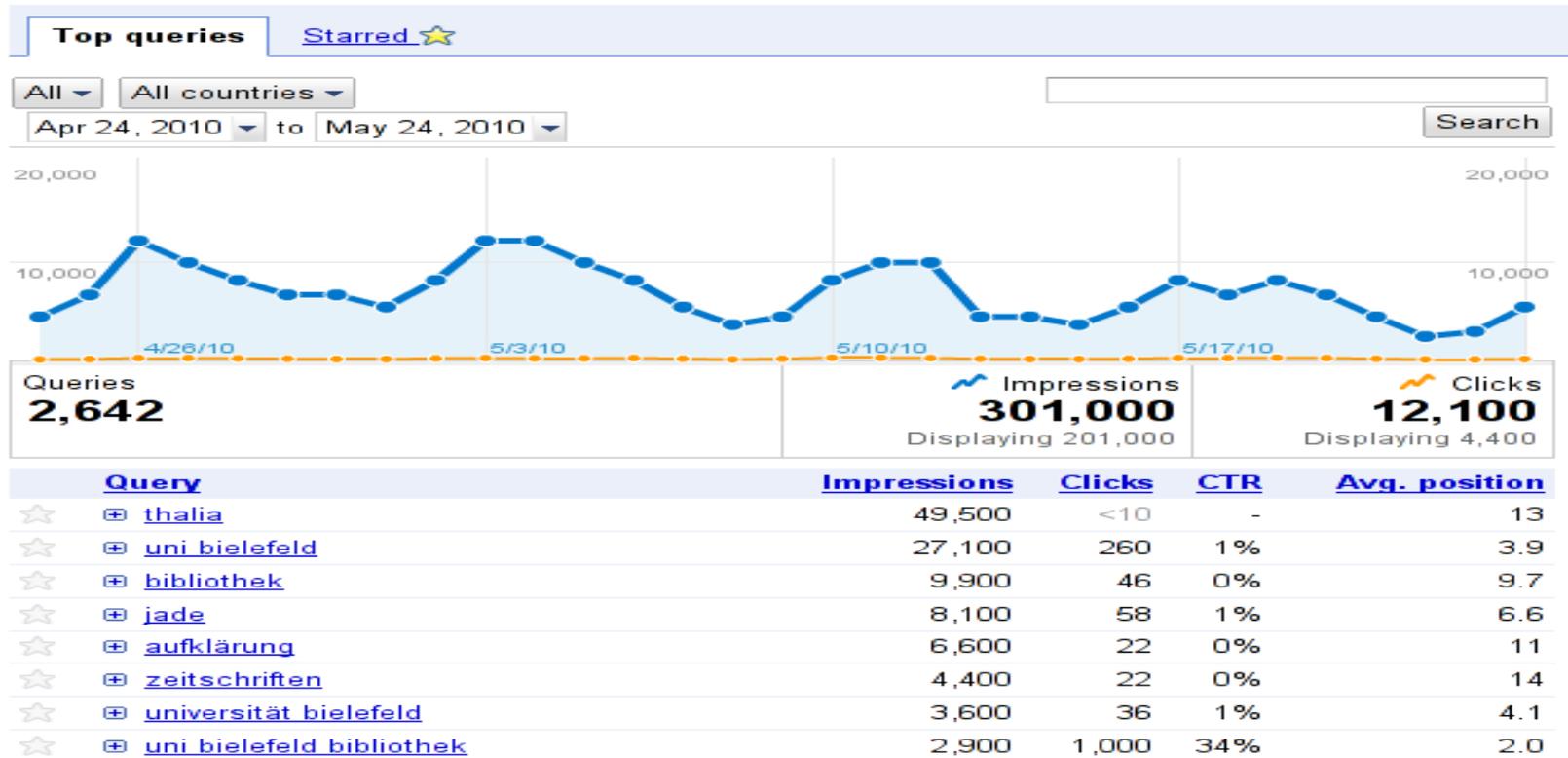
 <b>Seitwert</b>	 <b>Ranking (Top 100)</b>	
<p><b>54,88</b> von 100 Punkten</p>	<p><b>393</b> von 541135 Seiten</p>	

## Seitwert-Berechnung

	Gewichtung bei Google:	<b>19,05 / 29</b>		<b>66%</b>	<a href="#">Details ▼</a>
	Gewichtung bei Yahoo:	<b>8,14 / 17</b>		<b>48%</b>	<a href="#">Details ▼</a>
	Externe Wertungen:	<b>6,34 / 10</b>		<b>63%</b>	<a href="#">Details ▼</a>
	Technische Details:	<b>12 / 13</b>		<b>92%</b>	<a href="#">Details ▼</a>
	Social Bookmarks:	<b>0,35 / 22</b>		<b>2%</b>	<a href="#">Details ▼</a>
...	Sonstiges:	<b>9 / 9</b>		<b>100%</b>	<a href="#">Details ▼</a>

# Suchmaschinen-Optimierung (SEO)

## Search queries



# Ausblick

- Relevanzbewertung verstärkt nach Eigenschaften und Verhalten des Nutzers („Local Search“)
- „Universal Search“ = Neben Ergebnissen aus Webseiten auch Anzeige von Bildern, News, Blogbeiträgen etc.
- Weitere Fortschritte bei der semantischen Suche

Vielen Dank für Ihre  
Aufmerksamkeit!

[Präsentation beenden](#)