

# Maschinelle Verarbeitung von Daten im Netz verstehen

## Suchmaschinen und Datenbanksysteme

Skript zum Workshop auf der Herbstlichen InformatikLehrerFortbildung

Rostock, 20.9.2017

Lutz Hellmig



CC BY-NC-SA 3.0 DE

„Wo ist mein Puzzleteil?“ sah ich vor geraumer Zeit eine Grundschülerin vergebens in die Eingabemaske einer Suchmaschine eintippen. Einerseits war die Enttäuschung groß, auf diesem Wege nicht wieder an das verlorene Teil zu gelangen – andererseits ergab sich der Anlass für die Frage, auf welchem Wege Suchmaschinen Antworten auf Suchanfragen liefern.

„Denn wer da hat, dem wird gegeben“ (Matthäus 25, 29).



Das Wissen darum, wie Daten maschinell erhoben, verarbeitet und ausgewertet werden können, ist für die zielgerichtete Suche nach Informationen und bei der Bewertung des Suchergebnisses nützlich.

Einige für den Unterricht relevante Fragestellungen sollen an dieser Stelle besprochen werden.

## Welche Suchtechniken gibt es?

Intuitiv werden Anfragen an Suchmaschinen einerseits teilweise durch die Angabe eines oder mehrerer Suchbegriffe gestellt – die Autoergänzung der Suchbegriffe mit weiteren semantisch verwandten Begriffen liefert entsprechende Impulse. Andererseits werden auch häufig komplett ausformulierte Fragen in die Suchmaske eingegeben. Auch mit dieser Methode können für bestimmte Probleme durch die verbesserte Analyse der Suchanfragen Antworten auf die Fragen erzielt werden, wenngleich nicht immer mit derselben hohen Ausbeute

knoblauch eis

Alle Shopping Bilder Videos Maps Mehr Einstellungen Tools

Ungefähr 623.000 Ergebnisse (0,45 Sekunden)

Hinweise zum Datenschutz bei Google

SPÄTER ERINNERN ANSEHEN

Knoblaucheis - Rezept - ichkoche.at  
<https://www.ichkoche.at/knoblaucheis-rezept-69833>  
Gelatine in Wasser einweichen. Milch kurz aufwallen lassen, Zucker und Salz darin verflüssigen, die Gelatine dazumischen und verflüssigen.

Knoblauch eis Rezepte | Chefkoch.de  
[www.chefkoch.de/rs/s0/knoblauch+eis/Rezepte.html](https://www.chefkoch.de/rs/s0/knoblauch+eis/Rezepte.html)  
34 beliebte knoblauch eis Rezepte auf Chefkoch.de - Deutschlands bester Rezepteseite.

Bier- und Knoblauch-Eis schmeckt nur wenigen - derwesten.de  
<https://www.derwesten.de/Staete/Ennepetal>  
12.08.2012 - Schwelm. Wie wäre es mal mit Knoblauch-Eis? Oder Bier-Eis? Gibt's nicht? Zugegeben: Kassenschlager wären solche Eissorten wohl nicht.

Knoblaucheis Rezept | Webkoch.de  
<https://www.webkoch.de/rezept/knoblaucheis-22220>  
Das Rezept Knoblaucheis schnell mit Schritt für Schritt Anleitung nachkochen! ✓ ✓ Das Rezept wurde von 1 Benutzern mit durchschnittlich gut bewertet.

Bilder zu knoblauch eis

→ Weitere Bilder zu knoblauch eis

Unangemessene Bilder melden

Ich brauche ein Rezept für Knoblaucheis

Alle Shopping Bilder Videos Maps Mehr Einstellungen Tools

Ungefähr 4.490 Ergebnisse (0,51 Sekunden)

Hinweise zum Datenschutz bei Google

SPÄTER ERINNERN ANSEHEN

Knoblauch eis Rezepte | Chefkoch.de  
[www.chefkoch.de/rs/s0/knoblauch+eis/Rezepte.html](https://www.chefkoch.de/rs/s0/knoblauch+eis/Rezepte.html)  
34 beliebte knoblauch eis Rezepte auf Chefkoch.de - Deutschlands bester Rezepteseite.

Knoblaucheis - Rezept - ichkoche.at  
<https://www.ichkoche.at/knoblaucheis-rezept-69833>  
Gelatine in Wasser einweichen. Milch kurz aufwallen lassen, Zucker und Salz darin verflüssigen, die Gelatine dazumischen und verflüssigen.

Knoblaucheis Rezept | Webkoch.de  
<https://www.webkoch.de/rezept/knoblaucheis-22220>  
Das Rezept Knoblaucheis schnell mit Schritt für Schritt Anleitung nachkochen! ✓ ✓ Das Rezept wurde von 1 Benutzern mit durchschnittlich gut bewertet.

Schwarzes Sesameis - Carl Tode Göttingen  
<https://carltode.de/rezepte/schwarzes-sesameis/>  
17.07.2016 - Rezept für schwarzes Sesameis, Carl Tode Göttingen, Eis, Dessert.; ... Eissorten – Spinat, Knoblaucheis oder Eis mit Hafischflosse?

Kochen mit Fernet-Branca: Roman  
<https://books.google.de/books?isbn=3608937609>  
James Hamilton-Paterson - 2005  
Knoblaucheis mit Fernet-Branca hat vielleicht keine besondere Finesse, aber es ist hochwirksam, und mit dem Rezept, will mir scheinen, habe ich Ihnen ...

Abbildung 1: Schlagwortanfrage vs. ausformulierte Anfrage

Für eine verfeinerte Suche sollten die Schüler auch auf die Vielzahl an Möglichkeiten hingewiesen werden, die die Option der Erweiterten Suche bei den Suchmaschinen bietet. Zu den Optionen gehören die Suche nach einer genauen Formulierung (Phrasensuche) durch das Setzen von Anführungszeichen. Die Optionen sind zum Teil vom Medientyp (Text, Bild, Video, ...) abhängig: Bei der Bildsuche beispielsweise gibt es Filter, die die Ergebnisse nach der zugrundeliegenden Lizenz, dem bestimmenden Farbton oder der Bildgröße auswählen. Mittlerweile sind Rückwärtssuchen von Bildern möglich: Über die Suchmaske der Bildersuche werden Bilder hochgeladen und danach Entsprechungen im Netz gesucht. Nicht zuletzt kann die Ausbeute der Suchanfragen durch die Verwendung von Spezialsuchmaschinen verbessert werden. Prominente Beispiele sind Kindersuchmaschinen (Blinde Kuh, Frag Finn) oder Google Scholar, eine Suchmaschine für wissenschaftliche Artikel. Lohnenswert kann auch ein Wechsel der Suchmaschine sein. Metasuchmaschinen (Metager, Metacrawler, ...) übernehmen die simultane Befragung mehrerer Suchmaschinen.

### Wie werden Daten für die Suche aufbereitet?

Die Suchtechniken von Suchmaschinen lassen sich anhand von Kochbüchern anschaulich erklären. In Kochbüchern finden verschiedene Techniken Verwendung, mit denen das Auffinden von Informationen unterschiedlich gut erleichtert wird. Bei einer vergleichenden Betrachtung wird offensichtlich, dass ein Index die vorteilhafteste Möglichkeit ist, um Auffinden bestimmter Informationen zu ermöglichen.

Die Technik des Indizierens sollten die Schüler von Schulbüchern kennen. Im Weiteren ist die durch den Kurznachrichtendienst Twitter bekannte Technik des „Hashtags“ – sinntragende Worte werden mit einem vorangestellten Doppelkreuz # gekennzeichnet – aus informatischer Sicht ebenfalls nichts anderes als ein Indizieren einer Nachricht.

Die Suche auf lokalen Systemen kann ebenfalls durch das Indizieren der Dateien – das Durchsuchen der Dateien auf Schlüsselworte und das Verwalten derselben in einem speziellen Datenbanksystem verbessert werden.

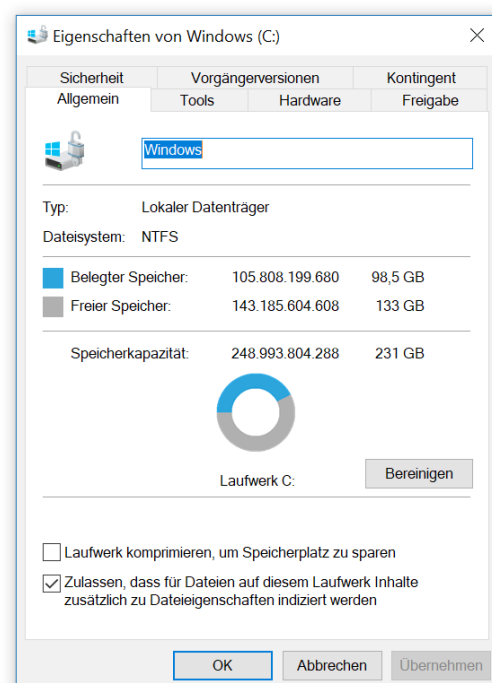


Abbildung 2: Laufwerke werden indiziert, um eine schnellere Suche zu ermöglichen.

Wie findet man effizient Informationen in Büchern?

Ein alphabetisches Rezeptregister ist für große Datenmengen unpraktisch.

Aioli	7	Knoblauchrahmsuppe mit Markklößchen	10
Blätterteig-Gyrostaschen	58	Knödel mit Rahmspinat, überbacken	24
Blätterteigtaler in Riesling-Käsesuppe	12	Lammkoteletts mit Brokkoli, gebratene	55
Bohneneintopf, französischer	14	Maccaroniauflauf, bunter	40
Camembert, eingelegter	71	Mangoldrouladen mit Schafskäsefüllung	30
Champignons in Gorgonzolasoße, gefüllte	22	Meeresfrüchtegratin in Knoblauchsoße	36
Chicorée aus dem Ofen	33	Mini-Mozzarella in Knoblauch-	
Coq au vin	51	Waldschlösschen-Fischsuppe mit Curry & Thymian	131
Crêpe mit Tomaten &		Weißkohleintopf	64
<b>Salate</b>		Wruckensuppe mit Tollatschen	23
Bücklingssalat	115	Zeesenboot-Aalsuppe	61
Feldsalat mit Pilzen	11	<b>Fisch</b>	
Heringssalat	164	Aufgerollte Seezunge	274
Kasseler-Broiler-Salat	123	Barschfilet mit Petersilienkruste	12
Matjessalat mit Apfel	32	Boddenzander auf Barther Tomaten-Carpaccio	57
Schwarzwurzel-Salat	107	Brathering oder Grüner Hering	145
Störtebeker-Salat	257	Dorsch in Folie mit Puszta-Salat	63
<b>Suppen</b>		Dorsch in Senf-Kruste	62
Borner Aalsuppe	147	Dorschrolade	127
Bouillabaisse mit einer Terrine von Bodden- & Ostseefischen	249	Eingelegte Heringe mit Tomaten	261

Die Einteilung der Rezepte in Kategorien erleichtert die zielgerichtete Suche bestimmter Gerichte.

In einem Index werden Begriffe gesammelt, inhaltlich zusammengefasst und die Orte des (ggf. mehrfachen) Vorkommens im Dokument festgehalten.

Synonyme werden durch Referenzen auf den Hauptbegriff (siehe ...) abgebildet.

## Index

Abpackdatum (siehe Haltbarkeit)	Diabetes mellitus (siehe Zuckerkrankheit)
Afrika 22	Diät (siehe Reisdiet)
- Reiszubereitung 131	Diätetik 40 f.
- Reisanbau 21 f.	Dunkler Reis (siehe auch Wilder Reis) 28 ff.
Ägypten 16, 23	Eisen 38, 39
Aleuronschicht 26 f.	Eiweiß 35, 39, 57
Alkoholische Getränke 28	- Kleber- 35
Amerika (siehe USA)	Entschlackungsdiät 81, 105 ff.
Anbau 21 f.	Entspelzen 26
Asien 15 ff.	Enzyme 35
Bali 25	Europa 15, 24
Ballaststoffe 26, 83	Fasten (siehe Reisfasten)
- und Darmkrebs 83	Fett 36, 39
Basmatireis 123	Flüssigkeitsausscheidung 37
Baustoffe (siehe auch Kohlenhydrate, Fette, - Eiweißstoffe, Vitamine) 33 ff.	Frankreich 16, 25
Bergreis (siehe Trockenreis)	Erbsen 68, 83 f.

Das Indexieren von Informationen ist aus technischer Sicht an den Umgang mit einem Datenbanksystem gebunden. Datenbanksysteme finden in vielfachen Kontexten Verwendung. Anhand der Indizierung durch Suchmaschinen lernen Schüler exemplarisch das grundsätzliche Prinzip von Datenbanksystemen kennen.

Das Herzstück eines Datenbanksystems ist eine Datenbank oder ein Verbund aus mehreren Datenbanken, in denen die Daten gespeichert werden. Zum Schutz der Daten und ihrer Struktur wird der Zugriff auf die Daten durch ein Datenbankmanagementsystem geregelt – einer Schnittstelle zwischen Benutzer und Datenbank.

In Kurzform:

**Datenbanksystem = Datenbank(en) + Datenbankmanagementsystem**

Eine Analogie zu Datenbanksystemen sind Bibliotheken. Die Bücherregale entsprechen den Datenbanken. Der Zugriff auf die Datenbanken wird durch die Bibliothekare (Datenbankmanagementsystem) geregelt: Sie überprüfen, wer Bücher ausleihen darf, sie helfen ggf. beim Auffinden von Büchern, sie allein stellen zurückgegebene Werke wieder in die Regale ein und sie sind auch für Neuanschaffungen und die Indexierung von Büchern verantwortlich. Man stelle sich vor, was passieren würde, wenn es keine Bibliothekare gäbe.

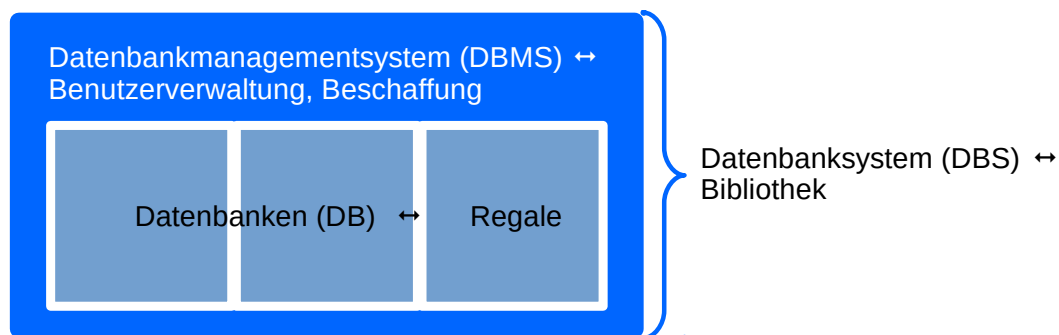


Abbildung 3: Analogie zwischen Datenbanksystemen und Bibliotheken

## Wie arbeiten Suchmaschinen?

Um Suchanfragen in kurzer Zeit beantworten zu können, werten Suchmaschinen den Index aus, den sie zuvor selbst erstellt haben.

**Soekia** ist die didaktische Simulation einer Suchmaschine, mit der die Arbeitsweise von Suchmaschinen anhand einer überschaubaren Anzahl kurzer Texte anschaulich experimentell erschlossen werden kann. Soekia ist als offline-Version und online über [www.soekia.ch](http://www.soekia.ch) erhältlich.

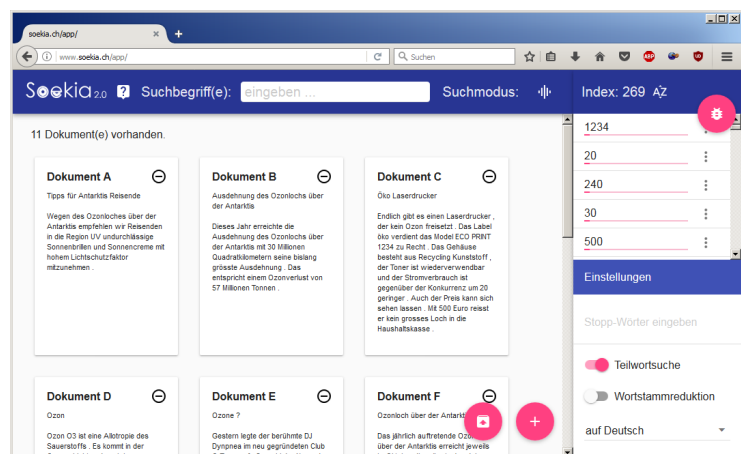


Abbildung 4: Eine Kollektion von Dokumenten in SOEKIA

Sowohl die Erstellung des Indexes als auch das Durchsuchen der Dokumente ist durch die Schüler mit SOEKIA erfahrbare. Mit Soekia können die Teilwortsuche (Wird „Ozonloch“ beim Suchbegriff „Ozon“ gefunden?) und die Wortstammreduktion (Wird das Wort „verbinden“ beim Suchbegriff „Verbindung“ gefunden?) manuell ein- bzw. ausgeschaltet werden. Der Einfluss der Sprache (Deutsch, Englisch, Französisch) auf die Wortstammreduktion kann ebenfalls beobachtet werden. Das Definieren von Stopp-Wörtern führt zur Nichtberücksichtigung von häufig verwendeten Füllwörtern (und, ist, ein, der, die, das, in, ...). Mit dem Klick auf den Schaltknopf „Crawler“ wird der Index jedesmal neu erstellt.



Abbildung 5: Crawler-Button

Die Suche erfolgt intuitiv. Das Vorkommen der Suchbegriffe in den durchsuchten Dokumenten wird visualisiert. Zu beachten ist, dass lediglich die Dokumente in Soekia durchsucht werden und nicht das gesamte Internet.

Hintergründe für die Funktionsweise von Soekia und Hinweise für den Unterrichtseinsatz sind auf <http://www.swisseduc.ch/informatik/soekia/materialien.html> zu finden. Die mit Soekia angebotenen Lernaufgaben für Schüler (<http://www.soekia.ch/AufgabenSoekia.pdf>) sind relativ anspruchsvoll und sollten für die Belange der Regionalschule deutlich reduziert werden. In jedem Fall sollte das Potential von Soekia genutzt werden, Vermutungen zum Verhalten von Suchmaschinen aufzustellen und experimentell zu prüfen. In einem weiteren Schritt können die Schüler untersuchen, welche der in Soekia ausprobierten Optionen durch reale Suchmaschinen genutzt werden.

## Wie finden Suchmaschinen Daten?

Der Klick auf den „Crawler“ in Soekia führte zum Durchsuchen der zur Verfügung stehenden Dokumente. Reale Suchmaschinen suchen auf ihnen bereits bekannten Seiten Links zu anderen Seiten und folgen ihnen. Auf diese Weise kann nur ein gewisser Teil der im Internet vorhandenen Ressourcen, aber nicht das gesamte Netz durchsucht werden. Der Teil des Netzes, der durch Suchmaschinen nicht gefunden wird, wird auch als „Deep Web“ bezeichnet.

## Wie finden Suchmaschinen genau die für den Sucher relevanten Ergebnisse?

Anhand von Soekia kann zwar die inhaltliche Auswertung von Dokumenten und das Ranking von Suchergebnissen nach objektiven Kriterien nachvollzogen werden, jedoch ist damit der Erfolg des fast weltweit uneingeschränkten Marktführers – Google – nicht erklärbar. Der Erfolg von Google resultierte unter anderem daraus, dass Google nicht nur die Inhalte der Seiten selbst untersucht, sondern auch Metadaten für die Rangierung der Suchergebnisse hinzuzieht.

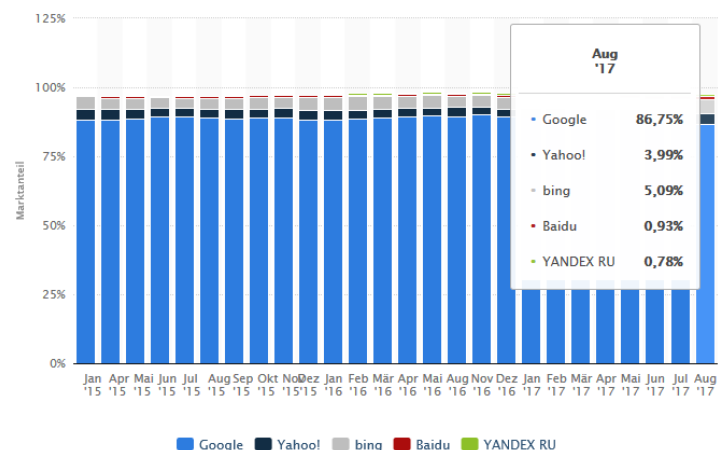


Abbildung 6: Marktanteile der Suchmaschinen auf Desktop-Systemen (Stand: 08/2017)

Metadaten sind beschreibende Attribute von Daten, anders gesagt: „Daten über Daten“. Metadaten gibt es zu Inhalten im Netz (Webseiten) als auch zu Akteuren im Netz (Benutzerverhalten).

Die Relevanz einer Webseite ist nicht allein aus dem Inhalt selbst ableitbar, sondern auch, wie viele andere Nutzer diese Seite für wichtig halten. Dies ist anhand der Zahl der Verweise zu ermitteln. Je höher die Zahl der Links auf eine Seite ist, umso größer ist die Wahrscheinlichkeit für eine hohe Relevanz der Seite. Noch bessere Ergebnisse erhält man, wenn die Relevanz der Seite, von der der Link ausgeht, in die Rangierung einbezogen wird. Ein Link von der Wikipedia dürfte zu einer höheren Rangierung führen als ein Link von einer privaten Hobby-Webseite mit wenigen Besuchern.

Zu den Metadaten, die das Benutzerverhalten charakterisieren, gehören Angaben zum Herkunftsort des Suchers, zur verwendeten Sprache und zu bisherigen Suchen und besuchten Webseiten. Eine besondere Rolle spielen in diesem Zusammenhang „Cookies“ – kleine Dateien, die auf dem Benutzerrechner angelegt werden und in denen Informationen zum Verhalten auf bisher besuchten Webseiten gespeichert werden.



Für die Untersuchung des Verhaltens der Suchmaschinen bieten sich vergleichende Betrachtungen an. Neben Google sollte auch ein Konkurrenzprodukt mit einer ähnlichen Philosophie (z. B. Bing) und eine Suchmaschine, die die Privatsphäre des Nutzers berücksichtigen (z. B. DuckDuckGo) getestet werden.

Vorschläge für entsprechende Arbeitsblätter liegen bei.

### **Was bedeutet das für mich?**

Der Komfort, auf die eigene Person zugeschnittene Suchergebnisse zu erhalten, wird mit der Preisgabe persönlicher Daten bezahlt. Diese sorgen dafür, dass kommerzielle Angebote zielgerichtet und individualisiert platziert werden können. Mithin besteht eine Zwickmühle zwischen Bequemlichkeit im Umgang mit Informatiksystemen einerseits und der Vertraulichkeit der eigenen Daten andererseits, derer die Schüler sich in ihren Handlungen bewusst sein sollten.

Durch die Berücksichtigung persönlicher Profile bei der Rangierung von Ergebnissen kann der Effekt der „Filterblase“ eintreten. Über die Recherche erhält man zunehmend Informationen, die der Erwartungshaltung des Benutzers entsprechen und ihn in seinem Weltbild bestärken. Dem entgegenstehende Informationen rangieren weiter hinten und dringen nicht mehr in das Bewusstsein des Suchenden. Dies verzerrt die Sicht auf die Welt. In ähnlicher Weise ist dieser Effekt bei Musik-Streaming-Diensten beobachtbar: Der Hörer erhält die Musik, die er mit ziemlicher Sicherheit gern hört – und nichts anderes. Oder würde Spotify einem Klassikliebhaber ein Heavy-Metal-Stück anbieten?

### **Nachtrag: Wolfram Alpha – die Antwortmaschine**

Mit Wolfram Alpha hat Steven Wolfram 2009 eine neue Klasse von Recherchewerkzeugen der Öffentlichkeit zugänglich gemacht. Wolfram Alpha ist keine reine Suchmaschine, sondern versucht, Antworten auf Suchanfragen mithilfe mathematischer und informatischer Methoden selbst zu generieren. Dazu beruft sich Wolfram Alpha unter anderem auf Datenbanken, die vorher durch Experten begutachtet wurden. Wenn man davon absieht, dass Wolfram Alpha nur für englischsprachige Suchanfragen optimiert ist, liefert Wolfram Alpha erstaunliche Ergebnisse.

### **Leseempfehlungen (Zuletzt geprüft am 26.9.2017)**

Historische Entwicklung und Marktanteile der Suchmaschinen

- <http://www.stuttgarter-zeitung.de/inhalt.entwicklung-der-suchmaschinen-und-am-anfang-war-nicht-google.1176b35d-592a-4e24-9a72-0c20fb7e9ec7.html>
- <https://de.statista.com/statistik/daten/studie/225953/umfrage/die-weltweit-meistgenutzten-suchmaschinen/>
- <https://seo-summary.de/suchmaschinen/>

Wolfram Alpha

- [https://de.wikipedia.org/wiki/Wolfram\\_Alpha](https://de.wikipedia.org/wiki/Wolfram_Alpha)