Google - Suchalgorithmen & Privacyprobleme

C. Bojko, D. Pape

8. Januar 2021

Übersicht

- Suchalgorithmus
- Privacyprobleme

Suchalgorithmus

Definition "How Google Search Works"- by Google

"Every time you search, there are thousands, sometimes millions, of webpages with helpful information. How Google figures out which results to show starts long before you even type, and is guided by a commitment to you to provide the best information."

Website "crawling"

Crawling

Der Crawling-Prozess beginnt mit einer Liste von Webadressen aus früheren Crawls und Sitemaps. Er besucht Websites, benützt Links zu neuen Websites und "springt" sozusagen immer weiter. Weitere Programme werten den Inhalt aus und verarbeiten die Daten.

Sitemaps

Eine Datei in der Informationen über die Seiten, Videos und anderen Dateien und deren zusammenhang stehen. Sie helfen der Google-Suche wichtige Informationen von der Website leichter zu finden.

Indexing

Indexing

Die im Crawling-Prozess gesammelte Information wird in einem gigantischen Index gesammelt, der auf sehr viele Rechner verteilt ist. Die Informationen sollen widerspiegeln, was das Thema einer Seite ist.

Ranking

Geschichte - PageRank

PageRank war der ursprüngliche Suchalgorithmus von Google in den 90ern. Die Neuheit von PageRank war, dass Seiten nach der Anzahl *Backlinks* gereiht wurden.

Backlinks

Eine Seite S hat n Backlinks, wenn es n Links zu S gibt.

In PageRank wurde jedem Backlink auch ein Gewicht zugewiesen - wenn die Seite, von der der Backlink stammt, selbst "wichtig"ist, dann zählt ein solcher Backlink mehr als ein Backlink von einer unwichtigen Seite.

Ranking - Probleme

Missbrauchspotenzial

Da Google akademische Ursprünge hat, war der PageRank Algorithmus öffentlich, und somit war es trivial, mit sog. Linkfarmen für den Algorithmus zu optimieren.

Teilprobleme des Rankings

Das Ziel ist, die *Relevanz* der Suchergebnisse für den Nutzer zu maximieren.

- Tippfehler
- Unterstützung von Natural Language Queries
- Mehrdeutige Begriffe
  - "how to **change** a light bulb"
  - "does post office **change** foreign currency"
  - "how to **change** laptop brightness"

Ranking - Probleme

Teilprobleme des Rankings

- Wird spezifische/aktuelle Information gesucht?
  - Eine Suche nach "bundesliga"oder "apple aktie"erfordert aktuelle Information
  - "burgerista öffnungszeiten"fordert einen spezifischen Datenpunkt
  - Ob aktuelle Information gesucht wird, wird u.A. daran festgestellt, ob die Such-Keywords *trending* sind.

Mit der Zeit wurden viele Updates zum Algorithmus eingeführt, die dieses Probleme beheben sollten.

Updates zum Suchalgorithmus

Panda (2011)

Bewertung der *Qualität* von Seiten:

- Ist die SEO sauber? (Metadaten, ...)
- Formulierung der Texte
- Qualität der Grafiken
- Sind die Unterseiten einander ähnlich?
- Wie lange bleiben die Besucher im Allgemeinen?

Penguin (2012)

Abwertung von Seiten, die "Black Hat "SEO einsetzen:

- Texte künstlich mit übermäßig vielen Keywords anreichern
- für die Nutzer unsichtbare Texte in die Seite einzubauen, um mehr Keywords unterzubringen (z.B. mit white-on-white text)

Updates zum Suchalgorithmus

Hummingbird (2013)

Erlaubte Suchen nicht nur per Keyword, sondern per Semantik: z.B. könnte man statt "wm 2014 gewinner "nach "wer hat die wm 2014 gewonnen "suchen und auf sinnvolle Ergebnisse hoffen.

RankBrain (2015)

ML/AI für Ranking:

- Ziel: Bessere Bearbeitung von zuvor unbekannten Suchanfragen
  - 15% aller Anfragen wurden vorher noch nie gesehen!
- AI soll die Semantik der Anfrage analysieren
- So können ähnliche zuvor beantwortete Anfragen zur Hilfe gezogen werden

Ranking

Heutzutage ist der Suchalgorithmus also klüger geworden und versteht auch den Inhalt der Seiten. Damit ist Black Hat SEO weniger effektiv. Weiters ist die genaue Funktionsweise nicht mehr öffentlich. Bekannt ist aber, dass u.a. folgende Variablen berücksichtigt werden:

Berücksichtigte Variablen

- Standort
  - Eine Suche nach "bicycle repair shop"in Paris sollte ganz andere Ergebnisse liefern als eine in Tokio.
  - Allerdings sollte dies für eine Suche nach "last nobel prize winner"nicht der Fall sein.
- Datum und Zeit
  - Eine Suche nach "wm sieger"sollte den Sieger der *letzten* WM finden.
- Sprache

Ranking

Berücksichtigte Variablen

- Vorherige Suchanfragen
  - Liefern Kontext
  - Wiederholte Anfragen bedeuten, dass vorherige Suchen nicht das richtige Ergebnis geliefert haben
- Gerät des Nutzers
  - Hauptsächlich geht es um Desktop, Tablet oder Mobile
  - Beispielsweise werden mobile-friendly Seiten bevorzugt, wenn der Nutzer auf einem Mobilgerät ist.

Privacyprobleme

Was sind Privacyprobleme

Privacyprobleme konzentrieren sich auch die Offenlegung der genetischen Informationen eines Users an Dritte.

Google's privacy policy (March 1, 2012)

Seit 2012 kann Google Daten über eine Vielzahl von Diensten weitergeben. Dazu gehören Websites Dritter, die Adsense oder Google Analytics verwenden.

Eric Schmidt - Google CEO 2009

"If you have something that you don't want anyone to know, maybe you shouldn't be doing it in the first place.",

Cookies

Was sind Cookies

Ein Cookie ist eine Textinormation, die im Browser eines Endgerätes zu einer Besuchten Website gespeichert werden kann. Dies kann entweder vom Webserver an der Browser oder direkt im Browser mit einem Javascript erstellt werden. Bei wiederholtem aufrufen einer Website kann der Server diese Datei auslesen.

Welche Daten werden gespeichert?

Cookies können eine Vielzahl von Informationen beinhalten. Unter anderem können persönliche Daten wie Name, Adresse, Email oder Telefonnummer gespeichert werden. Jedoch hat eine Website nur zu den Cookies zugriff die Sie selbst bereitstellen.

Cookies

Cookies bei Google

- Google platziert eins oder mehrere Cookies auf dein Rechner eines Users
- Dadurch kann sein Suchverlauf und der Besuch auf Websites gespeichert werden.
- Wenn ein Google Konto verknüpft ist werden alle besuchten Websites und alle Suchen dazu gespeichert.
- Seit 2016 informieren Googles Datenschutzrichtlinien nicht, ob und wann solche Aufzeichnungen gelöscht werden.
- Diese Informationen werden, auf Anfrage, Strafverfolgungsbehörden weitergeleitet.

Tracking

Wie verwendet Google Tracking?

Tracking kann durch verschiedene Tools von Google genutzt werden. Unter anderem durch Analytics, Play Services, reCAPTCHA, Google Fonts und APIs. Durch diese Tools kann Google die Route, die ein User nützt ermitteln. Durch die große Anzahl von Diensten und Websiten die zu Google gehören, ist es sehr schwer nachzuvollziehen wo die Informationen hingehen.

Ein Trick von Google

Als Beispiel gibt es das ReCAPTCHA von Google. Dieses Tool verwendet als Domaine "www.google.com"was bedeutet sie können Cookies die von Google stammen benützen und erweitern, da sie die dadurch die Beschränkung, dass man Drittanbieter Cookies verändert, umgehen.

Google Mail

Wie wird Google Mail wirklich verwendet

- Google Mail wird rein für Marketing Zwecke benützt.
- Mail's werden zwar nicht von Menschen gelesen, jedoch werden sie von Computern analysiert.
- Dadurch kann ein User genauestens analysiert werden, da in den Mails meist private Gespräche vorhanden sind.
  - EMails von Bestellungen (Amazon etc)
  - Private Mails mit Freunden/Bekannten

Zitat Eric Schmidt Google CEO 2010

"We know where you are. We know where you've been. We can more or less know what you're thinking about."

Google Mail

Third-Party Apps

2014 hat Google third-party Entwicklern Zugang zur GMail API gegeben. Dadurch können sie Software erstellen, die innerhalb der Plattform verwendet werden kann.

- Meist sind solche Softwares Produktivitäts-Tools.
  - Aufgabenmanager
  - Apps zum Signieren von Dokumenten
- Das Problem besteht darin, dass solche Apps beim installieren auch die Erlaubnis zum Lesen von Emails bekommen.

Third-Party Entwickler

Nach einem Beitrag im Wall Street Journal im Jahr 2018 gab es einen Vorfall, wo ein Angestellter solcher Drittanbieter 8000 Emails durchgelesen haben um ihre Algorithmen zu verbessern.

Google Chrome

Google's Suchleiste

- Chrome verfügt über eine kombinierte Such-/Adresszeile die auch Omnibar genannt wird.
- Google überträgt dabei in Echtzeit um die Sucheingabe zu vervollständigen.
- Der Gründer von Fluid New Media, Ahad Bokhari, fand durch das Testen des Browsers im Debug-Proxy Mode heraus, dass bei fast jedem Tastenschlag mit dem Server kommuniziert wird.
- Diese Daten kombiniert mit Cookies/Emails und Verlauf können ein sehr detailliertes Bild eines Users erstellen.

Android

Android ist das mit rund 72% am meisten genützte Smartphone OS. Es gibt rund 2,5 Milliarden aktive Android Geräte.
Beispiel:

- The Associated Press hat herausgefunden, dass auch wenn man Standortinformationen deaktiviert, der Standort im Google Konto hinterlegt wird. Wenn auch etwas ungenauer.
- Eine App kann den Bildschirmstatus des Smartphones ohne jegliche Berechtigung überwachen.
- Eine App kann Ihre WiFi-/Mobilfunk-Datennutzung ohne jegliche Berechtigungen überwachen.
- Eine App kann eine Liste aller anderen auf Ihrem Telefon installierten Apps ohne jegliche Berechtigungen erhalten.
- Apps können ohne explizite Berechtigung abfragen in welchem Winkel und in welche Richtung das Telefon gehalten wird und ob es sich bewegt.

Statistiken

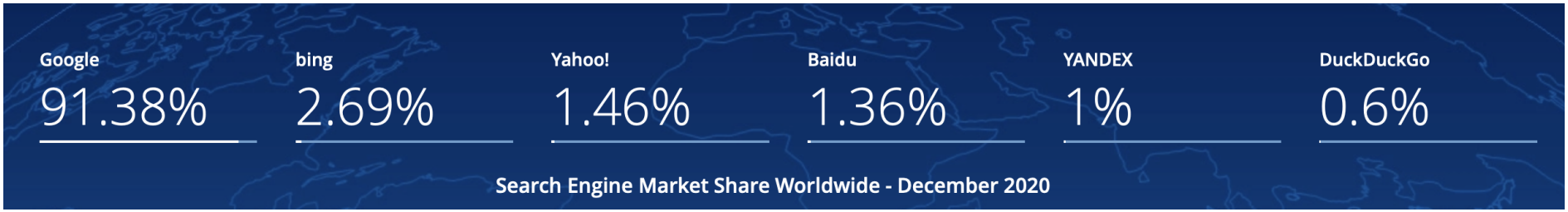

Abbildung: https://gs.statcounter.com/search-engine-market-share

Abbildung: https://gs.statcounter.com/browser-market-share

Statistiken

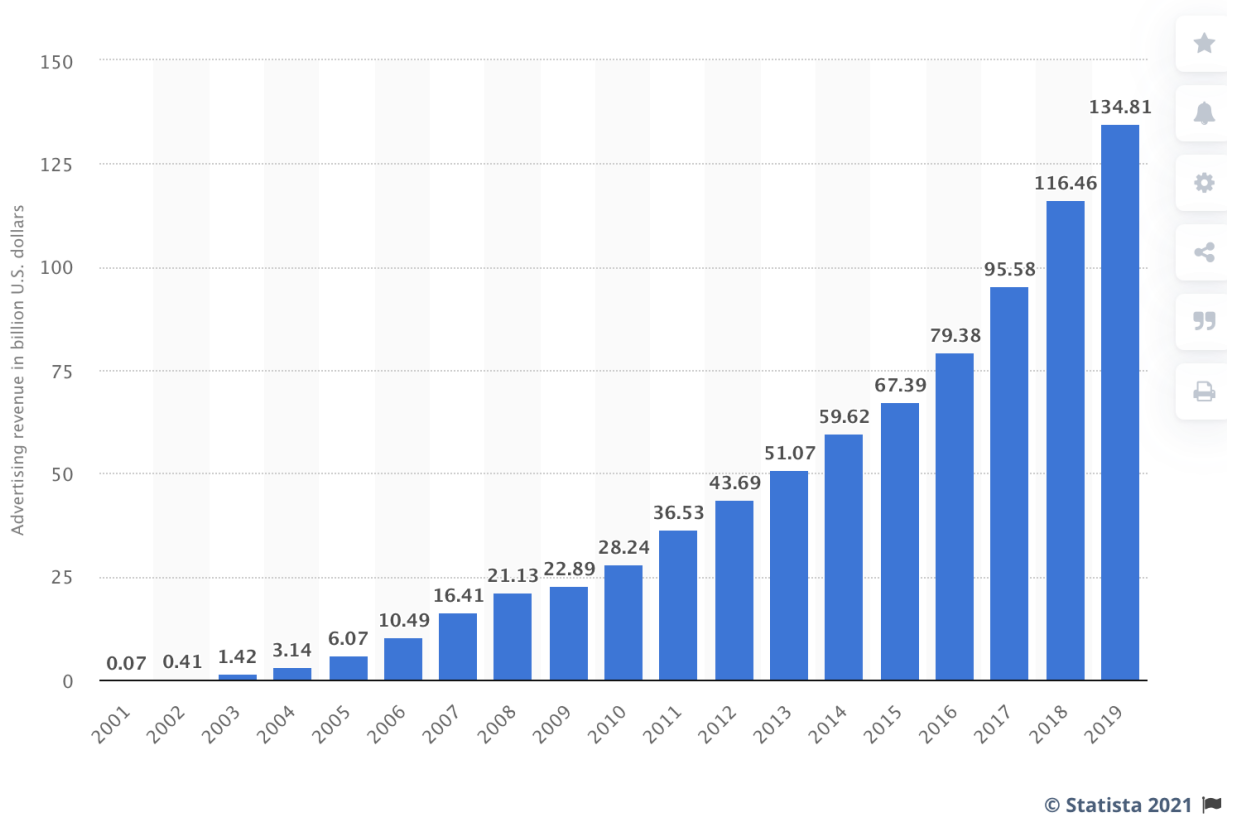

Abbildung: https://www.statista.com/statistics/266249/advertising-revenue-of-google/

Statistiken

Abbildung: https://emailclientmarketshare.com/

Abbildung: https://gs.statcounter.com/os-market-share/mobile/worldwide

Referenzen

- https://developers.google.com/search/docs/basics/how-search-works
- https://www.heise-regioconcept.de/google/wie-funktioniert-der-algorithmus-von-google
- https://www.google.com/search/howsearchworks/algorithms/
- https://protonmail.com/blog/google-privacy-problem/
- https://en.wikipedia.org/wiki/Privacy_concerns_regarding_Google
- https://www.wsj.com/articles/techs-dirty-secret-the-app-developers-sifting-through-your-gmail-1530544442
- https://smallbusiness.chron.com/google-chrome-privacy-problems-28257.html
- https://www.theverge.com/2019/5/7/18528297/google-io-2019-android-devices-play-store-total-number-statistic-keynote
- https://gs.statcounter.com/os-market-share/mobile/worldwide
- https://gs.statcounter.com/browser-market-share
- https://emailclientmarketshare.com/
- https://github.com/databurn-in/Android-Privacy-Issues