Engineer

Søkerobot

🕷 et program som henter sider for å indeksere dem

Hva det er

En søkemotor må selv gå ut og finne sidene på nettet, fordi ingen fører en samlet liste over dem, og programmet den sender, er en søkerobot. Den starter fra adresser den allerede kjenner, henter hver av dem over HTTP, leser lenkene i det som kommer tilbake, og legger de nye adressene til på listen sin. Googles veiledning om hvordan søket fungerer, gir de andre navnene som er i bruk, «also known as a crawler, robot, bot, or spider», altså også kjent som crawler, robot, bot eller edderkopp. RFC 9309, standarden for å holde roboter ute, beskriver søkeroboter som automatiske klienter søkemotorer kjører «to recursively traverse links for indexing», for å følge lenker rekursivt med tanke på indeksering.

Lenker er hovedveien inn. Google skriver at Googlebot finner nye adresser «primarily from links embedded in previously crawled pages», først og fremst fra lenker i sider som allerede er gjennomsøkt, og veiledningen gir det kjente eksempelet med en kategoriside som lenker til et nytt innlegg. Et sitemap er den andre veien, en liste eieren av nettstedet leverer direkte. En side som står utenfor begge veiene, gir en søkerobot ingen mulighet til å oppdage at den finnes.

Hentingen er det første av tre trinn Google beskriver. Gjennomsøkingen (*crawling*) laster ned tekst, bilder og video fra en side, indekseringen analyserer det som ble lastet ned, og lagrer det, og visningen (*serving*) besvarer et søk fra den indeksen. Under gjennomsøkingen gjengir Google også siden og kjører JavaScript-koden i en nyere versjon av Chrome, fordi Google uten det trinnet «might not see that content», kanskje ikke ser det innholdet. Alle trinnene skjer uten garanti: Google «doesn’t guarantee that it will crawl, index, or serve your page», garanterer altså ikke at siden blir gjennomsøkt, indeksert eller vist, og «doesn’t accept payment to crawl a site more frequently», tar ikke imot betaling for å gjennomsøke et nettsted oftere.

En søkerobot ventes også å si hvem den er. RFC 9309 lar hver søkerobot sette sitt eget navn, et produkttoken (*product token*) som bare består av bokstaver, understreker og bindestreker. Tokenet bør stå i identifikasjonsstrengen roboten sender, som over HTTP er headeren User-Agent, og strengen bør beskrive hva roboten skal gjøre. RFC-ens eksempel på en header inneholder `ExampleBot/0.1` og en lenke til en side om roboten, og den tilsvarende linjen i robots.txt lyder `user-agent: ExampleBot` . Navnet er en påstand og ikke en legitimasjon, siden hvilket som helst program kan sende det.

Google sender også automatiske forespørsler som faller utenfor gjennomsøkingen. Oversikten deres sorterer klientene i tre slag: vanlige søkeroboter som Googlebot, som «always respect robots.txt rules for automatic crawls», altså alltid følger reglene i robots.txt ved automatisk gjennomsøking, søkeroboter for spesialtilfeller, som arbeider under en avtale med nettstedet og kan følge de reglene eller la være, og hentere som en person utløser (*user-triggered fetchers*), som vanligvis gjør én enkelt forespørsel fordi noen ba om det, og som overser robots.txt.

Bruk begrepet i en beslutning

Bruk «Hva er en søkerobot?» i en konkret beslutning. Vi kan knytte begrepet til systemet dere bygger eller kjøper. Et nettsted eller en app som AI-assistentene kundene dine bruker, kan finne, lese og handle ut fra, kontrollert ved hver utgivelse i stedet for lovet én gang.

Se tjenestene våre Ta kontakt →

Undersøk antakelsene bak ordet

Bruk forklaringen og det tilknyttede arbeidet til å formulere spørsmål for prosjektet. Vi kan undersøke antakelsene sammen før et verktøy velges. Bygde et flerspråklig Hugo-nettsted på 3 språk — 108 malfiler som betjener tre språk og fire formater

Bygde et flerspråklig Hugo-nettsted på 3 språk Ta kontakt →

Hvorfor det er viktig

En søkerobot bestemmer selv når den kommer tilbake, og oppførselen dens avgjør en del av prisen for å bli funnet. Google skriver at Googlebot på de fleste nettsteder «shouldn't access your site more than once every few seconds on average», i snitt ikke bør besøke nettstedet oftere enn én gang med noen sekunders mellomrom, og søkerobotene trekker seg tilbake når et nettsted sliter. Med ordene i søkeveiledningen betyr HTTP 500-feil «slow down», altså senk farten. Googles veiledning om gjennomsøkingsbudsjett (*crawl budget*) beskriver hele mekanismen, men den er skrevet for nettsteder med omtrent en million sider som endres ukentlig, eller ti tusen som endres daglig, og Google sier at et nettsted der sidene gjennomsøkes samme dag som de publiseres, kan se bort fra den.

En søkerobot leser dessuten bare en viss mengde av hver fil. Når Googlebot gjennomsøker for Search, tar den de første 2 MB av en støttet filtype og de første 64 MB av en PDF, målt før komprimering, og overser resten.

Det svake punktet er navnet. Google advarer om at headeren User-Agent som Googlebot sender, «is often spoofed by other crawlers», ofte blir forfalsket av andre søkeroboter, så en logglinje som sier Googlebot, beviser ingenting i seg selv. Google dokumenterer to sjekker. For én enkelt adresse gjør du et omvendt DNS-oppslag, bekrefter at vertsnavnet slutter på `googlebot.com` , `google.com` eller `googleusercontent.com` , slår så opp det navnet forover og bekrefter at det gir den samme adressen. Googles eget eksempel:

```
host 66.249.66.1
1.66.249.66.in-addr.arpa domain name pointer crawl-66-249-66-1.googlebot.com.
host crawl-66-249-66-1.googlebot.com
crawl-66-249-66-1.googlebot.com has address 66.249.66.1
```

For mange adresser på en gang sammenligner du hver av dem med IP-områdene Google publiserer som JSON-filer i CIDR-notasjon, én fil per type klient. Begge sjekkene bygger på nettverket en forespørsel kom fra, og det kan ikke avsenderen bare skrive inn i en header.

Slik ser det ut i praksis

Serveren til dette nettstedet fører en tilgangslogg, og personvernsiden forteller hvordan søkeroboter skilles ut i den. Hver linje beholder nettleserens beskrivelse av seg selv, «user agent», «som er måten vi skiller mennesker fra søkemotorenes søkeroboter på», og et trinn til følger: «En søkerobot som oppgir et navn, sjekkes mot nettverkene eieren publiserer, ved hjelp av den tilslørte adressen nedenfor.» Det er Googles andre metode, og den passer til det loggen tar vare på. Den siste delen av hver IP-adresse erstattes med nuller før linjen når disken, så den fullstendige adressen et omvendt oppslag trenger, blir aldri lagret.

Sammendragene behandler søkeroboter som et eget publikum: «det samme for alt som er automatisert» ved siden av de tretti dagene som gjelder for mennesker, en daglig opptelling av «så mange sider lest, så mange av dem av søkeroboter, så mange som nådde kontaktsiden», og en månedlig rangering av

«hvilke søkeroboter og automatiske skannere som var innom». En søkerobot som oppgir navnet sitt ærlig, telles under sitt eget navn, og et kjent navn sendt fra et nettverk eieren ikke publiserer, stryker på sjekken.

Siden om å være klar for AI-agenter tar utgangspunkt i den samme leseren: en agent begynner der en søkerobot begynner, ved robots.txt-filen, sitemapet og søkeindeksene. Svaret siden gir, er å gjøre hvert av dem til en bevisst beslutning, helt ned til en IndexNow-melding ved hver produksjonssetting, slik at søkemotorene som er med, «hører om en ny side samme dag som den publiseres, og ikke først ved neste besøk».

Datoer er én måte et nettsted snakker med en søkerobot på, og de kan ende opp med å si lite som er til nytte. Omtalen av et sitemap som ga 172 av de 176 sidene sine ett tidsstempel sier rett ut hva som står på spill: «en endringsdato er et løfte til en søkerobot om at innholdet på en side har endret seg». Et nettsted som gir det løftet for alle sidene på en gang, fortsetter den, «forteller enten sannheten om en fullstendig omskriving eller forteller ingen noe nyttig».

Hvor det går galt

Forespørsler som kaller seg Googlebot, kommer fra noen andre enn Google. Loggen viser mye trafikk under navnet til Googlebot, og eieren slipper den enten inn som en velkommen gjest eller blokkerer navnet helt. Siden andre søkeroboter ofte kopierer det navnet, er begge reaksjonene risikable: den første stoler på en bedrager, og den andre stenger den ekte søkeroboten ute sammen med den. Sjekk nettverket forespørselen kom fra, med omvendt og forover DNS for én adresse eller mot Googles publiserte områder for mange, og bestem deg ut fra det i stedet for ut fra navnet.

En publisert side dukker aldri opp i gjennomsøkingen. Siden har ingen lenker som peker til seg, eller den kan bare nås etter innlogging. En søkerobot lærer om en adresse fra lenker og sitemaps, siden et sentralt register ikke finnes, og Googles veiledning nevner sider bak innlogging blant dem søkeroboten ikke kommer til. Lenk til siden fra et sted som allerede blir gjennomsøkt, før den opp i sitemapet, og hold alt som er ment for søk, utenfor innloggingen.

Gjennomsøkingstakten faller etter en dårlig periode for serveren. Færre sider hentes hver dag, og endringer tar lengre tid å bli sett. Trege svar, 5xx-feil og 429-svar senker alle Googles grense for gjennomsøkingskapasitet, som bare klatrer tilbake når svarene blir stabile og raske, og alle Googles søkeroboter på verten

deler det som er igjen. Rett feilene og latensen først. På et nettsted med noen få tusen sider er Googles veiledning om gjennomsøkingsbudsjett feil sted å begynne.

En side uten lenker dukker opp i forespørslene fra en søkerobot. Adressen ble aldri publisert, kanskje et privat utkast eller en testkopi, og den står likevel i tilgangsloggen under navnet til en søkerobot. Google kaller det «almost impossible to keep a site secret by not publishing links to it», nesten umulig å holde et nettsted hemmelig ved å la være å publisere lenker til det, fordi et klikk fra den siden til et annet nettsted kan legge igjen adressen i en henvisningslogg (*referrer log*) som blir publisert. Legg alt som er privat, bak passordbeskyttelse, som Google nevner som måten å holde både søkeroboter og mennesker ute på.

Les mer

- RFC 9309: Robots Exclusion Protocol — standardens syn på søkeroboter: produkttokens, strengen i User-Agent og reglene en søkerobot blir bedt om å respektere.
- In-depth guide to how Google Search works — oppdagelse, gjennomsøking, gjengivelse og indeksering slik Google beskriver dem.
- Overview of Google crawlers and fetchers (user agents) — de tre slagene klienter hos Google, grensen for filstørrelse, protokoller og caching.
- Verify requests from Google crawlers and fetchers — DNS-oppslagene og de publiserte IP-områdene, med gjennomgåtte eksempler.
- Optimize your crawl budget — hvordan Google setter gjennomsøkingstakten, og hvilke nettsteder som trenger å justere den.
- Googlebot — søkerobotene for mobil og datamaskin, hvor ofte Googlebot kommer innom, grensen for filstørrelse og hvorfor en side uten lenker ikke er hemmelig.

Ordliste

Velg ut fra deres faktiske rammer

Beskriv arbeidsflyten, eksisterende grensesnitt og brukerne som trenger endringen. Vi kan definere en testbar leveranse og en overlevering som gjør teamet i stand til å vedlikeholde løsningen.

Se hva det koster Ta kontakt →

Gjør begrepene praktiske

Spør hvordan «Hva er en søkerobot?» gjelder i organisasjonen deres. Et konkret eksempel fra dagens system er nok til å starte samtalen.

Se tjenestene våre

welcome@engineer.company +45 50 10 61 84 København, Danmark

Ta med deg denne siden

Visittkort

Laget med kjærlighet av D.

CVR 45969703 · Tagensvej 105C, 3. th, 2200 København N

Opphavsrett © 2025 – nå · Engineer ApS · Alle rettigheter forbeholdt.

Skrevet ut fra https://engineer.company/nb/glossary/crawler/