Engineer

Sökrobot

🕷 ett program som hämtar sidor för att indexera dem

Vad det är

En sökrobot är ett program som på egen hand hämtar sidor och följer länkarna i dem, oftast för att en sökmotor ska kunna bygga ett index. Det finns ingen central förteckning över webbens sidor, så sökmotorn måste själv leta upp dem: roboten börjar med adresser den redan känner till, hämtar var och en över HTTP och lägger till de nya adresser som länkarna i svaret pekar på. På engelska heter den crawler, och Googles guide till hur sökningen fungerar räknar upp de andra namnen: robot, bot och spider, alltså spindel. RFC 9309, standarden för Robots Exclusion Protocol, beskriver sökrobotar som automatiska klienter som sökmotorer kör ”to recursively traverse links for indexing”, för att rekursivt följa länkar inför indexering.

Länkar är huvudvägen in. Google skriver att Googlebot hittar nya adresser ”primarily from links embedded in previously crawled pages”, främst genom länkar i sidor som redan har genomsökts, och guiden ger det välbekanta exemplet med en kategorisida som länkar till ett nytt inlägg. En webbplatskarta är den andra vägen, en lista som webbplatsens ägare lämnar över direkt. En sida utan någotdera ger en sökrobot inget sätt att få veta att den finns.

Hämtningen är det första av tre steg som Google beskriver. Genomsökningen laddar ned en sidas text, bilder och video; indexeringen analyserar det som laddades ned och lagrar det; visningen besvarar en sökfråga ur det indexet. Under genomsökningen renderar Google också sidan och kör dess JavaScript i en aktuell version av Chrome, eftersom Google utan det steget ”might not see that content”, kanske inte ser innehållet. Inget av stegen är utlovat: Google ”doesn’t guarantee that it will crawl, index, or serve your page”, garanterar alltså inte att sidan genomsöks, indexeras eller visas, och tar inte heller betalt för att genomsöka en webbplats oftare.

En sökrobot förväntas också tala om vem den är. RFC 9309 låter varje sökrobot sätta ett eget namn, en produkttoken som bara består av bokstäver, understreck och bindestreck, och säger att den token bör finnas med i den identifieringssträng som roboten skickar, vilket över HTTP är headern User-Agent, och att strängen bör beskriva vad roboten är till för. RFC:ns exempelheader bär `ExampleBot/0.1` och en länk till en sida om sökroboten, och den motsvarande raden i robots.txt lyder `user-agent: ExampleBot`. Namnet är ett påstående och inget bevis, eftersom vilket program som helst kan skicka det.

Alla automatiska förfrågningar från Google är inte genomsökningar. Googles översikt delar in dess klienter i tre slag: vanliga sökrobotar som Googlebot, som alltid följer reglerna i robots.txt vid automatisk genomsökning; specialrobotar, som arbetar under en överenskommelse med webbplatsen och kanske följer de reglerna och kanske inte; och hämtare som startas av användare, som oftast gör en enda förfrågan för att en person bad om den och som bortser från robots.txt.

Använd begreppet i ett beslut

Använd ”Vad är en sökrobot?” i ett konkret beslut. Vi kan knyta begreppet till systemet ni bygger eller köper. En webbplats eller app som de AI-assistenter dina kunder använder kan hitta, läsa och agera på, kontrollerad vid varje release i stället för utlovad en gång.

Se våra tjänster Hör av er →

Granska antagandena bakom ordet

Använd förklaringen och det länkade arbetet för att formulera frågor till projektet. Vi kan undersöka antagandena tillsammans innan ett verktyg väljs. Byggde en flerspråkig Hugo-webbplats på 3 språk — 108 mallfiler som täcker tre språk och fyra format

Byggde en flerspråkig Hugo-webbplats på 3 språk Hör av er →

Varför det spelar roll

En sökrobot bestämmer själv när den kommer tillbaka, och dess vanor avgör en del av vad det kostar att bli hittad. Google skriver att Googlebot på de flesta webbplatser i genomsnitt inte bör besöka webbplatsen oftare än en gång med några sekunders mellanrum, och dess sökrobotar backar när en webbplats har det svårt; med guidens ord betyder HTTP-fel 500 ”slow down”, sakta ner. Googles guide om genomsökningsbudget beskriver hela mekanismen, men den är skriven för webbplatser med ungefär en miljon sidor som ändras varje vecka eller 10 000 som ändras varje dag, och Google säger att en webbplats vars sidor genomsöks samma dag som de publiceras inte behöver den.

En sökrobot läser också bara en viss mängd av varje fil. Googlebot tar vid genomsökning för Google Sök de första 2 MB av en filtyp som stöds och de första 64 MB av en PDF, mätt före komprimering, och bortser från resten.

Den svaga punkten är namnet. Google varnar för att den User-Agent-header som Googlebot skickar ”is often spoofed by other crawlers”, ofta förfalskas av andra sökrobotar, så en loggrad som säger Googlebot bevisar ingenting i sig. Google dokumenterar två kontroller. För en enskild adress gör man ett omvänt DNS-uppslag, kontrollerar att värdnamnet slutar på `googlebot.com` , `google.com` eller `googleusercontent.com` , slår sedan upp namnet framlänges och kontrollerar att det ger samma adress tillbaka. Googles eget exempel:

```
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
```

För många adresser på en gång jämför man var och en med de IP-intervall som Google publicerar som JSON-filer i CIDR-notation, en fil per slag av klient. Båda kontrollerna vilar på nätverket som förfrågan kom från, och det kan avsändaren inte bara skriva in i en header.

Så ser det ut i praktiken

Den här webbplatsens server för en åtkomstlogg, och sidan om integritet säger hur sökrobotar skiljs från andra besökare i den. Varje rad behåller webbläsarens beskrivning av sig själv, ”som är hur vi skiljer människor från sökmotorernas sökrobotar”, och ett andra steg följer: ”En sökrobot som uppger ett namn kontrolleras mot de nätverk som dess ägare publicerar, med hjälp av den suddade adressen nedan.” Det är Googles andra metod, den som passar loggen. Den sista delen av varje IP-adress ersätts med nollor innan raden når disken, så den fullständiga adress som ett omvänt uppslag behöver lagras aldrig.

Sammanställningarna behandlar sökrobotar som en egen publik: ”detsamma för allt automatiserat” bredvid de trettio dagar som sparas för människor, en daglig sammanräkning av ”så många sidor lästa, så många av dem av sökrobotar, så många som nådde kontaktsidan” och en månatlig rangordning av ”vilka sökrobotar och automatiska skannrar som anropade”. En sökrobot som uppger sitt namn ärligt räknas under det namnet, och ett känt namn som skickas från ett nätverk som ägaren inte publicerar klarar inte kontrollen.

Sidan om att vara redo för AI-agenter utgår från samma läsare: ”En agent börjar där en sökrobot gör”, i robots.txt-filen, webbplatskartan och sökindexen. Svaret där är att göra var och en av dem avsiktlig, ända ned till ett meddelande via IndexNow vid varje driftsättning, så att de sökmotorer som deltar ”får veta om en ny sida samma dag som den publiceras i stället för vid sitt nästa besök”.

Datum är ett av sätten som en webbplats talar med en sökrobot på, och de kan sluta med att inte säga något användbart alls. Redogörelsen för en webbplatskarta som gav 172 av sina 176 sidor samma tidsstämpel säger rakt ut vad som står på spel: ”ett ändringsdatum är ett löfte till en sökrobot om att en sidas innehåll har ändrats”. En webbplats som ger det löftet 172 gånger samtidigt, fortsätter den, ”säger antingen sanningen om en fullständig omskrivning eller säger ingenting användbart till någon”.

Var det går fel

Förfrågningar som kallar sig Googlebot kommer från någon annan än Google. Loggen visar tung trafik under Googlebots namn, och ägaren släpper antingen igenom den som en välkommen gäst eller blockerar namnet rakt av. Eftersom andra sökrobotar ofta kopierar det namnet är båda reaktionerna osäkra: den första litar på en bedragare, och den andra stänger ute den äkta sökroboten tillsammans med den. Kontrollera källnätverket, med omvänd och framlänges DNS för en adress eller mot Googles publicerade intervall för många, och låt det avgöra, inte namnet.

En publicerad sida dyker aldrig upp i en genomsökning. Den har inga inkommande länkar, eller så går den bara att nå efter inloggning. Eftersom det inte finns något centralt register får en sökrobot veta om en adress genom länkar och webbplatskartor, och Googles guide räknar sidor bakom en inloggning till dem som dess sökrobot inte når. Länka sidan från något som redan genomsöks, lista den i webbplatskartan och håll allt som är tänkt för sökning utanför inloggningen.

Genomsökningstakten sjunker efter en dålig period för servern. Antalet sidor som hämtas per dag minskar, och ändringar tar längre tid att upptäcka. Långsamma svar, 5xx-fel och svar med 429 sänker alla Googles gräns för genomsökningskapacitet, som bara stiger igen när svaren blir stabila och snabba, och varje sökrobot från Google på samma värddator delar på det som blir över. Åtgärda felen och latensen först; på en webbplats med några tusen sidor är Googles guide om genomsökningsbudget inte rätt ställe att börja på.

En sida som aldrig har länkats dyker upp i en sökrobots förfrågningar. Adressen, kanske ett privat utkast eller en testkopia, publicerades aldrig men syns ändå i åtkomstloggen. Google kallar det ”almost impossible to keep a site secret by not publishing links to it”, nästan omöjligt att hålla en webbplats hemlig genom att låta bli att publicera länkar till den, eftersom ett klick från sidan till en annan webbplats kan lämna dess adress i en referrerlogg som sedan publiceras. Lägg allt privat bakom lösenordsskydd, som Google anger som sättet att hålla både sökrobotar och människor ute.

Läs mer

- RFC 9309: Robots Exclusion Protocol — standardens syn på sökrobotar: produkttokens, strängen i User-Agent och de regler som en sökrobot ombeds respektera.
- In-depth guide to how Google Search works — upptäckt, genomsökning, rendering och indexering så som Google beskriver dem.
- Overview of Google crawlers and fetchers (user agents) — Googles tre slag av klienter, gränsen för filstorlek, protokoll och cachning.
- Verify requests from Google crawlers and fetchers — DNS-uppslagen och de publicerade IP-intervallen, med genomräknade exempel.
- Optimize your crawl budget — hur Google sätter sin genomsökningstakt, och vilka webbplatser som behöver justera den.
- Googlebot — sökrobotarna för mobil och dator, hur ofta Googlebot kommer på besök, gränsen för filstorlek och varför en sida utan länkar inte är hemlig.

Ordlista

Välj utifrån era faktiska förutsättningar

Beskriv arbetsflödet, befintliga gränssnitt och användarna som behöver förändringen. Vi kan definiera en testbar leverans och en överlämning som gör att teamet kan underhålla lösningen.

Se vad det kostar Hör av er →

Gör begreppen praktiska

Fråga hur ”Vad är en sökrobot?” gäller i er organisation. Ett konkret exempel från ert nuvarande system räcker för att börja samtalet.

Se våra tjänster

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

Ta med dig den här sidan

Visitkort

Omsorgsfullt skapad med kärlek av D.

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

Upphovsrätt © 2025 – nu · Engineer ApS · Alla rättigheter förbehållna.

Utskrivet från https://engineer.company/sv/glossary/crawler/