Engineer

Crawler

🕷 et program, der henter sider for at indeksere dem

Hvad det er

Der findes ingen samlet liste over nettets sider, så en søgemaskine må selv ud og finde dem, og det program, den sender, er en crawler. Den begynder med adresser, den allerede kender, henter hver af dem over HTTP, læser linkene i svaret og føjer de nye adresser til sin liste. Googles guide til, hvordan Search virker, nævner de andre navne, der er i brug, nemlig robot, bot og spider, og RFC 9309, standarden for robots.txt, beskriver crawlere som automatiske klienter, som søgemaskiner bruger til at følge links rekursivt med henblik på indeksering.

Links er hovedvejen ind. Ifølge Google finder Googlebot nye adresser først og fremmest via links i sider, den allerede har crawlet, og guiden giver det velkendte eksempel med en kategoriside, der linker til et nyt indlæg. Et sitemap er den anden vej, en liste, som ejeren af sitet afleverer direkte. En side uden nogen af delene giver en crawler ingen måde at opdage, at den findes.

At hente er det første af tre trin, Google beskriver. Crawling henter tekst, billeder og video fra en side; indeksering analyserer det hentede og gemmer det; visning besvarer en søgning ud fra indekset. Under crawlingen renderer Google også siden og kører dens JavaScript i en nyere udgave af Chrome, fordi Google uden det trin risikerer at overse indhold. Intet af det er lovet: Google garanterer ikke, at en side bliver crawlet, indekseret eller vist, og tager ikke imod betaling for at crawle et site oftere.

En crawler forventes også at sige, hvem den er. RFC 9309 lader hver crawler vælge sit eget navn, et produkttoken, der kun må bestå af bogstaver, understregninger og bindestreger, og siger, at tokenet bør indgå i den identifikationsstreng, crawleren sender, hvilket over HTTP er User-Agent-headeren, og at strengen bør beskrive crawlerens formål. RFC'ens eksempel på

en header indeholder `ExampleBot/0.1` og et link til en side om crawleren, og den tilsvarende linje i robots.txt lyder `user-agent: ExampleBot` . Navnet er en påstand og ikke et bevis, for ethvert program kan sende det.

Ikke enhver automatisk forespørgsel fra Google er en crawl. Googles oversigt deler klienterne i tre slags: almindelige crawlere som Googlebot, der altid respekterer robots.txt ved automatisk crawling; specialcrawlere, der arbejder efter en aftale mellem sitet og et Google-produkt og måske følger reglerne, måske ikke; og brugerudløste fetchere, der som regel henter én side, fordi en person bad om den, og som ignorerer robots.txt.

Brug begrebet i en beslutning

Brug »Hvad er en crawler?« i en konkret beslutning. Vi kan knytte begrebet til det system, I bygger eller køber. Et site eller en app, som de AI-assistenter, jeres kunder bruger, kan finde, læse og handle på, kontrolleret ved hver udgivelse i stedet for lovet én gang.

Se vores ydelser Kontakt os →

Undersøg antagelserne bag ordet

Brug forklaringen og det tilknyttede arbejde til at formulere spørgsmål til jeres projekt. Vi kan undersøge antagelserne sammen, før et værktøj vælges. Har bygget et Hugo-site på 3 sprog — 108 skabelonfiler, der betjener tre sprog og fire formater

Har bygget et Hugo-site på 3 sprog Kontakt os →

Hvorfor det betyder noget

En crawler bestemmer selv, hvornår den kommer igen, og dens manerer afgør en del af prisen for at blive fundet. Google skriver, at Googlebot på de fleste sites i gennemsnit højst bør komme forbi én gang hvert par sekunder, og at crawlerne trækker sig, når et site har det svært; i Search-guiden betyder en HTTP 500-fejl, at de skal sætte farten ned. Googles guide til crawlbudget beskriver hele mekanismen, men den er skrevet til sites med omkring en million sider, der ændres ugentligt, eller ti tusind, der ændres dagligt, og Google siger, at et site, hvis sider bliver crawlet samme dag, de udgives, ikke behøver den.

En crawler læser også kun så meget af hver fil. Når Googlebot crawler til Search, henter den de første 2 MB af en understøttet filtype og de første 64 MB af en PDF, målt før komprimering, og ser bort fra resten.

Det svage punkt er navnet. Google advarer om, at andre crawlere ofte forfalsker den User-Agent-header, Googlebot sender, så en loglinje, der siger Googlebot, beviser intet i sig selv. Google beskriver to tjek. For en enkelt adresse laver man

et omvendt DNS-opslag, bekræfter, at værtsnavnet ender på `googlebot.com`, `google.com` eller `googleusercontent.com`, slår så navnet op den anden vej og bekræfter, at det giver den samme adresse. 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 én gang sammenligner man hver af dem med de IP-intervaller, Google udgiver som JSON-filer i CIDR-notation, én fil pr. slags klient. Begge tjek bygger på det netværk, forespørgslen kom fra, og det kan afsenderen ikke bare skrive ind i en header.

Sådan ser det ud i praksis

Siden om privatliv fortæller, hvordan crawlere skelnes fra andre i dette sites adgangslog. Hver linje gemmer browserens beskrivelse af sig selv, »som er sådan, vi skelner mennesker fra søgemaskinernes robotter«, og et trin mere følger: »En robot, der giver et navn, tjekkes mod de netværk, dens ejer offentliggør, med den slørede adresse nedenfor.« Det er Googles anden metode, og den passer til loggen: den sidste del af hver IP-adresse erstattes med nuller, før linjen rammer disken, så den fulde adresse, som et omvendt opslag kræver, gemmes aldrig.

Opgørelserne behandler crawlere som et publikum for sig: »det samme for alt det automatiske« ved siden af de tredive dage for mennesker, en daglig optælling af »så mange sider læst, så mange af dem af robotter, så mange der nåede kontaktsiden« og en månedlig rangliste over, »hvilke robotter og automatiske scannere der ringede på«. En crawler, der ærligt siger sit navn, tælles under det navn, og et kendt navn sendt fra et netværk, som ejeren ikke offentliggør, består ikke tjekket.

Siden om at være klar til AI-agenter tager udgangspunkt i den samme læser: »En agent begynder, hvor en crawler gør«, i robots.txt-filen, i sitemappet og i søgeindeksene. Dens svar er at gøre hvert af dem til et bevidst valg, helt ned til en IndexNow-melding ved hver udrulning, så de søgemaskiner, der er med, hører om en ny side »samme dag, den går i luften, og ikke først ved deres næste besøg«.

Datoer er én måde, et site taler til en crawler på, og de kan ende med ikke at sige noget brugbart. Beretningen om et sitemap, der gav 172 af dets 176 sider samme tidsstempel siger det ligeud: »en ændringsdato er et løfte til en crawler om, at en sides indhold er ændret«. Et site, der giver løftet for hver side på én gang, fortsætter den, »fortæller enten sandheden om en fuld omskrivning eller fortæller ingen noget brugbart«.

Hvor det går galt

Forespørgsler, der kalder sig Googlebot, kommer ikke fra Google. Loggen viser meget trafik under Googlebots navn, som ejeren enten lukker ind eller blokerer helt. Da andre crawlere ofte låner navnet, er ingen af reaktionerne sikker: den første stoler på en bedrager, og den anden lukker den ægte crawler ude sammen med den. Tjek netværket, som forespørgslen kom fra, med omvendt og fremadrettet DNS for én adresse eller mod Googles udgivne intervaller for mange, og afgør sagen ud fra det frem for navnet.

En udgivet side dukker aldrig op i en crawl. Intet linker til den, eller den kan kun nås efter login. Uden et centralt register lærer en crawler om en adresse via links og sitemaps, og Googles guide nævner sider bag et login blandt dem, crawleren ikke kan nå. Link til siden fra et sted, der allerede bliver crawlet, sæt den i sitemappet, og hold alt, der skal kunne findes i søgning, uden for login.

Crawlhastigheden falder efter en dårlig periode for serveren. Der hentes færre sider om dagen, og ændringer er længere tid om at blive set. Langsomme svar, 5xx-fejl og 429-svar sænker alle Googles grænse for crawlkapacitet, som først stiger igen, når svarene bliver stabile og hurtige, og alle Googles crawlere på værten deler det, der er tilbage. Ret fejlene og svartiderne først; et site med et par tusind sider skal ikke begynde i guiden til crawlbudget.

En side, som ingen har linket til, dukker op i en crawlers forespørgsler. Adressen blev aldrig udgivet, måske et privat udkast, og alligevel står den i adgangsloggen. Google kalder det næsten umuligt at holde et site hemmeligt ved bare at lade være med at linke til det, fordi et klik fra siden videre til et andet site kan efterlade adressen i en henvisningslog, der bliver offentliggjort. Læg alt privat bag adgangskodebeskyttelse, som Google nævner som måden at holde både crawlere og mennesker ude på.

Læs mere

- RFC 9309: Robots Exclusion Protocol — standardens syn på crawlere: produkttokens, User-Agent-strengen og de regler, en crawler bedes respektere.
- In-depth guide to how Google Search works — opdagelse, crawling, rendering og indeksering, sådan som Google beskriver dem.
- Overview of Google crawlers and fetchers (user agents) — Googles tre slags klienter, grænsen for filstørrelse, protokoller og caching.
- Verify requests from Google crawlers and fetchers — DNS-opslagene og de udgivne IP-intervaller med gennemregnede eksempler.
- Optimize your crawl budget — hvordan Google sætter sin crawlhastighed, og hvilke sites der har brug for at justere den.
- Googlebot — crawlerne til smartphone og computer, hvor ofte Googlebot kommer forbi, grænsen for filstørrelse, og hvorfor en side uden links ikke er hemmelig.

Ordliste

Vælg ud fra jeres faktiske vilkår

Beskriv arbejdsgangen, de eksisterende grænseflader og de brugere, der har brug for ændringen. Vi kan definere en testbar leverance og en overdragelse, der gør teamet i stand til at vedligeholde løsningen.

Se, hvad det koster Kontakt os →

Gør begreberne praktiske

Spørg, hvordan »Hvad er en crawler?« gælder i jeres organisation. Et konkret eksempel fra jeres nuværende system er nok til at begynde samtalen.

Se vores ydelser

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

Tag denne side med dig

Visitkort

Lavet med kærlighed af D.

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

Copyright © 2025 – nu · Engineer ApS · Alle rettigheder forbeholdes.

Udskrevet fra https://engineer.company/da/glossary/crawler/