Engineer

Краулер

🕷 програма, що завантажує сторінки для індексації

Що це таке

Спільного переліку сторінок мережі не існує, тож пошукова система мусить шукати їх сама, і програма, яку вона для цього посилає, — це краулер. Він починає з адрес, які вже знає, завантажує кожну через HTTP, читає посилання у відповіді й додає нові адреси до свого списку. Посібник Google про те, як працює пошук, наводить інші назви, що трапляються, — робот, бот, павук, а RFC 9309, стандарт robots.txt, описує краулери як автоматичних клієнтів, за допомогою яких пошукові системи рекурсивно проходять посиланнями для індексації.

Посилання — головний шлях усередину. За словами Google, Googlebot знаходить нові адреси передусім через посилання на сторінках, які вже просканував, і посібник наводить звичний приклад: сторінка категорії посилається на новий допис. Карта сайту — другий шлях, перелік, який власник сайту передає напряму. Сторінка без того й іншого не дає краулеру способу дізнатися, що вона існує.

Завантаження — перший із трьох етапів, які описує Google. Сканування завантажує текст, зображення й відео сторінки; індексація аналізує завантажене й зберігає його; показ відповідає на запит із цього індексу. Під час сканування Google також рендерить сторінку й виконує її JavaScript у свіжій версії Chrome, бо без цього кроку може не побачити частину вмісту. Нічого з цього не обіцяно: Google не гарантує, що просканує, проіндексує чи покаже сторінку, і не бере плати за те, щоб сканувати сайт частіше.

Краулер також має сказати, хто він. RFC 9309 дозволяє кожному краулеру обрати власне ім’я, токен продукту, що складається лише з літер, підкреслень і дефісів, і каже, що токен слід вказувати в ідентифікаційному рядку, який надсилає краулер, — через HTTP це заголовок User-Agent, — а

сам рядок має описувати призначення краулера. Приклад заголовка в RFC містить `ExampleBot/0.1` і посилання на сторінку про краулер, а відповідний рядок у robots.txt виглядає як `user-agent: ExampleBot`. Ім'я — це твердження, а не доказ, бо надіслати його може будь-яка програма.

Не кожен автоматичний запит від Google — це сканування. Огляд Google ділить клієнтів на три види: звичайні краулери на кшталт Googlebot, які завжди дотримуються robots.txt під час автоматичного сканування; спеціальні краулери, що працюють за домовленістю між сайтом і продуктом Google і можуть дотримуватися цих правил, а можуть і ні; та завантажувачі на запит користувача, які зазвичай роблять один запит, бо про це попросила людина, і ігнорують robots.txt.

Застосуйте поняття до рішення

Застосуйте «Що таке краулер?» до конкретного рішення. Ми можемо пов'язати поняття із системою, яку ви створюєте або купуєте. Сайт чи застосунок, який ШІ-асистенти ваших клієнтів можуть знайти, прочитати й використати, і це перевіряється з кожним випуском, а не обіцяється один раз.

Переглянути наші послуги Зв'язатися з нами →

Перевірте припущення за терміном

Використайте пояснення та пов'язані роботи, щоб сформулювати запитання для проєкту. Ми можемо перевірити припущення разом до вибору інструмента. Побудувала сайт на Hugo 3 мовами — 108 файлів шаблонів обслуговують три мови й чотири формати

Побудувала сайт на Hugo 3 мовами Зв'язатися з нами →

Чому це важливо

Краулер сам вирішує, коли повернутися, і його манери визначають частину ціни за те, щоб вас знаходили. Google пише, що на більшості сайтів Googlebot у середньому має звертатися не частіше ніж раз на кілька секунд і що його краулери відступають, коли сайтові важко; за посібником про пошук, помилка HTTP 500 означає, що треба сповільнитися. Посібник Google про бюджет сканування описує весь механізм, але написаний для сайтів приблизно з мільйоном сторінок, що змінюються щотижня, або з десятьма тисячами, що змінюються щодня, і Google каже, що сайту, сторінки якого сканують у день публікації, він не потрібен.

Краулер також читає лише частину кожного файлу. Googlebot, що сканує для пошуку, бере перші 2 МБ файлу підтримуваного типу й перші 64 МБ PDF, рахуючи до стиснення, а решту ігнорує.

Слабке місце — ім'я. Google попереджає, що заголовок User-Agent, який надсилає Googlebot, часто підробляють інші краулери, тож рядок журналу зі словом Googlebot сам по собі нічого не доводить. Google описує дві перевірки. Для однієї адреси виконайте зворотний DNS-запит, переконайтеся, що ім'я хоста закінчується на `googlebot.com`, `google.com` або `googleusercontent.com`, потім виконайте прямий запит для цього імені й переконайтеся, що він повертає ту саму адресу. Приклад самого Google:

```
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
```

Для багатьох адрес одразу звіряйте кожну з діапазонами IP, які Google публікує як файли JSON у нотації CIDR, по файлу на кожен вид клієнта. Обидві перевірки спираються на мережу, з якої прийшов запит, а її відправник не може просто вписати в заголовок.

Як це виглядає на практиці

Сторінка про приватність пояснює, як у журналі доступу цього сайту відрізняють краулери від інших відвідувачів. Кожен запис зберігає ідентифікаційний рядок браузера, «за яким ми відрізняємо людей від пошукових роботів», і далі йде ще один крок: «Якщо робот називає себе, ми звіряємо розмиту адресу (див. нижче) з мережами, які оприлюднює його власник.» Це другий метод Google, і саме він годиться для того, що зберігає журнал: остання частина кожної IP-адреси замінюється нулями ще до запису на диск, тож повна адреса, потрібна для зворотного запиту, не зберігається ніколи.

Зведення розглядають краулери як окрему аудиторію: «те саме для всього автоматизованого трафіку» поруч із тридцятьма днями для людей, щоденний підрахунок — «стільки-то сторінок прочитано, стільки-то з них — роботами, стільки-то дійшли до сторінки контактів» — і місячний рейтинг того, «які роботи та автоматичні сканери заходили на сайт». Краулер, що чесно себе називає, рахується під своїм ім'ям, а відоме ім'я, надіслане з мережі, якої власник не оприлюднює, перевірку не проходить.

Сторінка про готовність до ШІ-агентів виходить із того самого читача: «Агент починає там само, де краулер» — з файлу robots.txt, карти сайту і пошукових індексів. Її відповідь — зробити кожне з них свідомим вибором, аж до оголошення через IndexNow під час кожного розгортання, щоб пошуковики, які його підтримують, дізнавалися про нову сторінку «в день публікації, а не під час наступного обходу».

Дати — один зі способів, якими сайт говорить із краулером, і зрештою вони можуть не казати нічого корисного. Розповідь про карту сайту, що дала 172 зі 176 сторінок одну позначку часу прямо називає ставку: «дата зміни — це обіцянка пошуковому роботові, що вміст сторінки змінився». Сайт, який дає таку обіцянку для всіх сторінок одночасно, продовжує вона, «або правдиво повідомляє про повне переписування, або не повідомляє нікому нічого корисного».

Де виникають проблеми

Запити, що називають себе Googlebot, надходять не від Google. Журнал показує багато трафіку під іменем Googlebot, і власник або пускає його, або блокує це ім’я повністю. Оскільки інші краулери часто позичають це ім’я, обидві реакції ризиковані: перша довіряє самозванцеві, а друга разом із ним відрізає справжній краулер. Перевірте мережу, з якої прийшов запит, — зворотним і прямим DNS-запитом для однієї адреси або за опублікованими діапазонами Google для багатьох — і вирішуйте на цій підставі, а не за іменем.

Опублікована сторінка ніколи не з’являється в скануванні. На неї ніщо не посилається, або до неї можна дістатися лише після входу. Без центрального реєстру краулер дізнається про адресу з посилань і карт сайту, а посібник Google називає сторінки за входом серед тих, до яких його краулер дістатися не може. Поставте посилання на сторінку з місця, яке вже сканують, додайте її до карти сайту і тримайте все, що має бути в пошуку, поза входом.

Швидкість сканування падає після поганого періоду для сервера. Щодня завантажується менше сторінок, і зміни помічають пізніше. Повільні відповіді, помилки 5xx і відповіді 429 знижують ліміт пропускної здатності сканування Google, який знову зростає, лише коли відповіді стають стабільними й швидкими, а всі краулери Google на хості ділять те, що лишилося. Спершу виправте помилки й затримки; сайтові з кількома тисячами сторінок не варто починати з посібника про бюджет сканування.

Сторінка, на яку ніхто не посилався, з'являється в запитах краулера. Адресу ніколи не публікували, можливо, це приватна чернетка, і все ж вона є в журналі доступу. Google називає майже неможливим зберегти сайт у таємниці, просто не публікуючи посилань на нього, бо перехід із цієї сторінки на інший сайт може залишити її адресу в журналі переходів, який потім опублікують. Ховайте все приватне за захистом паролем — саме його Google називає способом не пускати ні краулери, ні людей.

Що почитати далі

- RFC 9309: Robots Exclusion Protocol — погляд стандарту на краулери: токени продукту, рядок User-Agent і правила, яких краулера просять дотримуватися.
- In-depth guide to how Google Search works — виявлення, сканування, рендеринг та індексація так, як їх описує Google.
- Overview of Google crawlers and fetchers (user agents) — три види клієнтів Google, ліміт розміру файлу, протоколи й кешування.
- Verify requests from Google crawlers and fetchers — DNS-запити й опубліковані діапазони IP із розібраними прикладами.
- Optimize your crawl budget — як Google встановлює швидкість сканування і яким сайтам її варто налаштовувати.
- Googlebot — краулери для смартфонів і комп'ютерів, як часто Googlebot заходить на сайт, ліміт розміру файлу і чому сторінка без посилань не є секретною.

Словник

Обирайте за фактичними умовами

Опишіть робочий процес, наявні інтерфейси та користувачів, яким потрібна зміна. Ми можемо визначити перевірюваний результат і спосіб передачі, який дасть вашій команді змогу підтримувати його.

Дивитися, скільки це коштує Зв'язатися з нами →

Зробіть термінологію практичною

Запитайте, як «Що таке краулер?» стосується вашої організації. Конкретного прикладу з поточної системи достатньо для початку розмови.

Переглянути наші послуги

welcome@engineer.company +45 50 10 61 84 Копенгаген, Данія

Візьміть цю сторінку із собою

Візитівка

Дбайливо зроблено з любов'ю — D.

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

Авторське право © 2025 – дотепер · Engineer ApS · Усі права захищено.

Надруковано з https://engineer.company/uk/glossary/crawler/