Engineer

rastreador web

🕷 descubrir páginas siguiendo enlaces y mapas del sitio

Qué es

No existe una lista maestra de las páginas de la web, así que un buscador tiene que salir a encontrarlas, y el programa que envía es un rastreador. Parte de direcciones que ya conoce, obtiene cada una por HTTP, lee los enlaces de la respuesta y añade las direcciones nuevas a su lista. La guía de Google sobre el funcionamiento de la Búsqueda recoge los otros nombres utilizados: «también conocido como crawler, robot, bot o spider». RFC 9309, el estándar de exclusión de robots, describe los rastreadores como clientes automatizados que los buscadores ejecutan «para recorrer enlaces de forma recursiva con fines de indexación».

Los enlaces son la principal vía de entrada. Google afirma que Googlebot encuentra direcciones nuevas «principalmente a partir de enlaces incluidos en páginas rastreadas anteriormente», y su guía ofrece el caso habitual de una página de categoría que enlaza una publicación nueva. Un mapa del sitio es la segunda vía: una lista que entrega directamente quien administra el sitio. Una página que no aparece en ninguna de las dos no ofrece al rastreador forma de saber que existe.

La obtención de páginas es la primera de las tres etapas que describe Google. El rastreo descarga el texto, las imágenes y los vídeos de una página; la indexación analiza lo descargado y lo almacena; la presentación de resultados responde a una búsqueda a partir de ese índice. Durante el rastreo, Google también renderiza la página y ejecuta su JavaScript en una versión reciente de Chrome, ya que sin ese paso «podría no ver ese contenido». Ninguna de las etapas está garantizada: Google «no garantiza que vaya a rastrear, indexar o mostrar su página» y «no acepta pagos para rastrear un sitio con más frecuencia».

También se espera que un rastreador diga quién es. RFC 9309 establece que cada uno elija su propio nombre, un identificador de producto formado únicamente por letras, guiones bajos y guiones, e indica que debería aparecer en la cadena de identificación que envía el rastreador, que por HTTP es la cabecera User-Agent, y que esa cadena debería describir su finalidad. La cabecera del ejemplo del RFC lleva `ExampleBot/0.1` y un enlace a una página sobre el rastreador, y la línea correspondiente de robots.txt dice `user-agent: ExampleBot` . El nombre es una declaración, no una credencial, porque cualquier programa puede enviarlo.

No toda petición automatizada de Google es un rastreo. Su descripción general clasifica los clientes en tres tipos: rastreadores comunes, como Googlebot, que «siempre respetan las reglas de robots.txt en los rastreos automáticos»; rastreadores de casos especiales, que trabajan conforme a un acuerdo con el sitio y pueden seguir o no esas reglas; y agentes de obtención activados por usuarios, que suelen hacer una sola petición porque alguien la ha solicitado e ignoran robots.txt.

Utilice el concepto en una decisión

Aplique «¿Qué es un rastreador web?» a una decisión real. Podemos ayudarle a relacionar el concepto con el sistema que está construyendo o comprando. Un sitio o una aplicación que los asistentes de IA de sus clientes puedan encontrar, leer y utilizar, con comprobaciones en cada versión en lugar de una promesa inicial.

Explore nuestros servicios Póngase en contacto →

Compruebe los supuestos detrás del término

Utilice la explicación y los trabajos enlazados para formular preguntas sobre su proyecto. Podemos examinar los supuestos con usted antes de elegir una herramienta. Sitio Hugo multilingüe en 3 idiomas — 108 archivos de plantillas para tres idiomas y cuatro formatos

Sitio Hugo multilingüe en 3 idiomas Póngase en contacto →

Por qué importa

Un rastreador decide por sí mismo cuándo volver, y su comportamiento determina parte del coste de que un sitio sea encontrado. Google dice que Googlebot «no debería acceder a su sitio más de una vez cada pocos segundos de media» en la mayoría de los sitios, y sus rastreadores reducen la actividad cuando un sitio tiene dificultades; en palabras de la guía de la Búsqueda, los errores HTTP 500 significan «reducir la velocidad». La guía de Google sobre el presupuesto de rastreo describe todo el mecanismo, pero está dirigida a sitios de aproximadamente un millón de páginas que cambian cada semana o diez mil que cambian a diario. Google indica que un sitio cuyas páginas se rastrean el día de su publicación no la necesita.

Un rastreador también limita cuánto lee de cada archivo. Al rastrear para la Búsqueda, Googlebot toma los primeros 2 MB de cualquier archivo de un tipo compatible y los primeros 64 MB de un PDF, medidos antes de la compresión, e ignora el resto.

El punto débil es el nombre. Google advierte que la cabecera User-Agent que envía Googlebot «suele ser falsificada por otros rastreadores», por lo que una línea del registro que diga Googlebot no demuestra nada por sí sola. Google documenta dos comprobaciones. Para una dirección, haga una consulta DNS inversa, confirme que el nombre de host termina en `googlebot.com`, `google.com` o `googleusercontent.com` y después resuelva ese nombre mediante una consulta directa para confirmar que devuelve la misma dirección. Este es el ejemplo de 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
```

Para muchas direcciones a la vez, compare cada una con los rangos IP que Google publica como archivos JSON en notación CIDR, un archivo por tipo de cliente. Las dos comprobaciones se basan en la red de la que procede la petición, algo que el remitente no puede limitarse a escribir en una cabecera.

Cómo se aplica en la práctica

El servidor de este sitio conserva un registro de accesos, y la página de privacidad explica cómo se distinguen en él los rastreadores. Cada línea conserva el agente de usuario, «que es lo que permite separar a las personas de los rastreadores de los buscadores», y después se aplica un segundo paso: «Un rastreador que da un nombre se comprueba frente a las redes que publica su propietario, utilizando la dirección enmascarada que se describe a continuación». Es el segundo método de Google y encaja con lo que conserva el registro. La última parte de cada dirección IP se sustituye por ceros antes de que la línea llegue al disco, de modo que nunca se almacena la dirección completa que necesitaría una consulta inversa.

Los resúmenes tratan a los rastreadores como un público propio: «lo mismo para todo lo automatizado» junto a los treinta días que se conservan para las personas; un recuento diario de «tantas páginas leídas, tantas por rastreadores, tantas visitas a la página de contacto»; y una clasificación mensual de «qué rastreadores y escáneres automatizados han pasado». Un rastreador que se identifica con honestidad se cuenta con su propio nombre, y un nombre famoso enviado desde una red que su propietario no publica no supera la comprobación.

La página sobre la preparación para agentes de IA parte del mismo lector: «Un agente empieza donde lo hace un rastreador», en el archivo robots.txt, el mapa del sitio y los índices de búsqueda. Su respuesta consiste en definir cada uno deliberadamente, hasta el anuncio mediante IndexNow en cada despliegue, para que los buscadores participantes «se enteren de una página nueva el día en que se publica, en lugar de en su próxima visita».

Las fechas son una forma de hablar con los rastreadores, y pueden acabar sin decir nada útil. El caso de un mapa del sitio que asignaba la misma marca de tiempo a 172 de sus 176 páginas expone claramente lo que está en juego: «una fecha de modificación es una promesa a un rastreador de que el contenido de una página ha cambiado». Un sitio que hace esa promesa para todas sus páginas a la vez, continúa, «o dice la verdad sobre una reescritura completa o no le dice nada útil a nadie».

Dónde surgen los problemas

Las peticiones que se llaman Googlebot no proceden de Google. El registro muestra mucho tráfico con el nombre de Googlebot, y quien administra el sitio lo deja pasar como un invitado bienvenido o bloquea el nombre por completo. Como otros rastreadores suelen copiarlo, ninguna reacción es segura: la primera confía en un impostor y la segunda excluye también al rastreador auténtico. Compruebe la red de origen mediante DNS inverso y directo para una dirección o frente a los rangos publicados por Google para muchas, y decida en función de ello, no del nombre.

Una página publicada nunca aparece en un rastreo. No hay ningún enlace hacia ella o solo se puede acceder después de iniciar sesión. Sin un registro central, un rastreador conoce las direcciones a través de enlaces y mapas del sitio, y la guía de Google incluye las páginas tras un inicio de sesión entre las que su rastreador

no puede alcanzar. Enlace la página desde algún lugar ya rastreado, inclúyala en el mapa del sitio y mantenga accesible sin iniciar sesión todo lo que deba aparecer en las búsquedas.

La frecuencia de rastreo baja después de una mala racha del servidor. Cada día se obtienen menos páginas y los cambios tardan más en detectarse. Las respuestas lentas, los errores 5xx y las respuestas 429 reducen el límite de capacidad de rastreo de Google, que solo vuelve a subir cuando las respuestas se estabilizan y son rápidas; todos los rastreadores de Google que visitan el host comparten la capacidad restante. Corrija primero los errores y la latencia; en un sitio de unos pocos miles de páginas, la guía de Google sobre el presupuesto de rastreo no es el punto de partida.

Una página sin enlaces entrantes aparece entre las peticiones de un rastreador. La dirección nunca se publicó, quizá porque era un borrador privado o una copia de prueba, y aun así aparece en el registro de accesos con el nombre de un rastreador. Google dice que es «casi imposible mantener un sitio en secreto solo por no publicar enlaces hacia él», ya que un clic desde esa página hacia otro sitio puede dejar su dirección en un registro de referencias que termine publicado. Proteja con contraseña todo lo privado, el método que Google señala para mantener fuera tanto a los rastreadores como a las personas.

Lecturas adicionales

- RFC 9309: Protocolo de exclusión de robots — los rastreadores según el estándar: identificadores de producto, cadena User-Agent y reglas que se les pide respetar.
- Guía detallada sobre el funcionamiento de la Búsqueda de Google — descubrimiento, rastreo, renderizado e indexación tal como los describe Google.
- Descripción general de los rastreadores y agentes de obtención de Google — los tres tipos de clientes de Google, el límite de tamaño de los archivos, los protocolos y la caché.
- Verificar peticiones de rastreadores y agentes de obtención de Google — las consultas DNS y los rangos IP publicados, con ejemplos desarrollados.
- Optimizar el presupuesto de rastreo — cómo fija Google la frecuencia de rastreo y qué sitios necesitan ajustarla.
- Googlebot — los rastreadores para teléfonos y ordenadores, la frecuencia de las visitas de Googlebot, su límite de tamaño de archivo y por qué una página sin enlaces entrantes no es secreta.

Glosario

Elija según sus restricciones reales

Describa el flujo de trabajo, las interfaces existentes y los usuarios que necesitan el cambio. Podemos definir una entrega comprobable y un traspaso que permita a su equipo mantener lo entregado.

Vea cuánto cuesta Póngase en contacto →

Haga práctica la terminología

Pregunte cómo se aplica «¿Qué es un rastreador web?» a su organización. Basta un ejemplo concreto de su sistema actual para iniciar la conversación.

Explore nuestros servicios

welcome@engineer.company +45 50 10 61 84 Copenhague, Dinamarca

Llévese esta página

Tarjeta de contacto

Creado con esmero y cariño por D.

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

Derechos de autor © 2025 – actualidad · Engineer ApS · Todos los derechos reservados.

Impreso desde https://engineer.company/es/glossary/crawler/