Engineer

Robot d'exploration

🕷 programme qui explore le web en suivant ses liens

Ce que c’est

Il n’existe pas de liste centrale des pages du web ; un moteur de recherche doit donc aller les trouver, et le programme qu’il envoie est un robot d’exploration. Il part d’adresses déjà connues, récupère chacune par HTTP, lit les liens de la réponse et ajoute les nouvelles adresses à sa liste. Le guide de Google sur le fonctionnement de la recherche donne les autres noms employés, « également appelé crawler, robot, bot ou spider », et la RFC 9309, norme d’exclusion des robots, décrit ces programmes comme des clients automatisés que les moteurs de recherche exécutent « pour parcourir récursivement les liens en vue de l’indexation ».

Les liens sont la voie principale. Google indique que Googlebot découvre de nouvelles adresses « principalement à partir de liens intégrés dans des pages déjà explorées », et son guide donne l’exemple familier d’une page de catégorie qui renvoie à un nouvel article. Un sitemap constitue la seconde voie, une liste que le propriétaire du site fournit directement. Une page sans l’un ni l’autre ne donne au robot aucun moyen d’apprendre son existence.

La récupération est la première des trois étapes décrites par Google. L’exploration télécharge le texte, les images et les vidéos d’une page ; l’indexation analyse ce qui a été téléchargé et le stocke ; la présentation des résultats répond à une recherche à partir de cet index. Pendant l’exploration, Google effectue aussi le rendu de la page en exécutant son JavaScript dans une version récente de Chrome, car sans cette étape il « pourrait ne pas voir ce contenu ». Aucune étape n’est garantie : Google « ne garantit pas qu’il explorera, indexera ou présentera votre page » et « n’accepte pas de paiement pour explorer un site plus fréquemment ».

Un robot est aussi censé s'identifier. La RFC 9309 prévoit que chacun définisse son nom, un jeton de produit composé uniquement de lettres, de traits de soulignement et de traits d'union. Elle indique que ce jeton devrait figurer dans la chaîne d'identification envoyée par le robot, qui est l'en-tête User-Agent en HTTP, et que cette chaîne devrait décrire l'objectif du robot. L'exemple d'en-tête de la RFC contient `ExampleBot/0.1` et un lien vers une page sur le robot ; la ligne robots.txt correspondante est `user-agent: ExampleBot` . Ce nom est une déclaration, pas un justificatif d'identité, puisque n'importe quel programme peut l'envoyer.

Les requêtes automatisées de Google ne relèvent pas toutes de l'exploration. Sa vue d'ensemble classe ses clients en trois catégories : les robots courants tels que Googlebot, qui « respectent toujours les règles de robots.txt pour les explorations automatiques » ; les robots spécialisés, qui travaillent dans le cadre d'un accord avec le site et peuvent respecter ces règles ou non ; et les outils de récupération déclenchés par un utilisateur, qui font généralement une seule requête à sa demande et ignorent robots.txt.

Utilisez le concept pour prendre une décision

Utilisez « Qu'est-ce qu'un robot d'exploration ? » pour éclairer une décision réelle. Nous pouvons vous aider à relier ce concept au système que vous construisez ou achetez. Un site ou une application que les assistants d'IA de vos clients peuvent trouver, lire et utiliser : une capacité vérifiée à chaque livraison, plutôt qu'une promesse ponctuelle.

Découvrir nos prestations Contactez-nous →

Vérifiez les hypothèses derrière le terme

Appuyez-vous sur l'explication et les réalisations liées pour formuler les questions de votre projet. Nous pouvons examiner les hypothèses avec vous avant le choix d'un outil. Site Hugo multilingue en 3 langues — 108 fichiers de modèles pour trois langues et quatre formats

Site Hugo multilingue en 3 langues Contactez-nous →

Pourquoi c'est important

Un robot décide lui-même quand revenir, et sa manière de se comporter détermine une partie du coût de la visibilité. Google indique que, sur la plupart des sites, Googlebot « ne devrait pas accéder au site plus d'une fois toutes les quelques secondes en moyenne », et ses robots ralentissent lorsqu'un site peine à répondre ; selon le guide de recherche, les erreurs HTTP 500 signifient « ralentissez ». Le guide du budget d'exploration décrit le mécanisme entier, mais s'adresse aux sites d'environ un million de pages modifiées chaque semaine ou dix mille modifiées chaque jour. Google précise qu'un site dont les pages sont explorées le jour de leur publication n'en a pas besoin.

Un robot ne lit par ailleurs qu'une quantité limitée de chaque fichier. Googlebot, lorsqu'il explore pour la recherche, prend les 2 premiers MB d'un fichier d'un type pris en charge, ou les 64 premiers MB dans le cas d'un PDF, tailles mesurées avant compression, puis ignore le reste.

Le point faible est le nom. Google avertit que l'en-tête User-Agent envoyé par Googlebot « est souvent usurpé par d'autres robots » ; une ligne de journal qui dit Googlebot ne prouve donc rien à elle seule. Google documente deux vérifications. Pour une seule adresse, effectuez une recherche DNS inverse, vérifiez que le nom d'hôte se termine par `googlebot.com` , `google.com` ou `googleusercontent.com` , puis effectuez une résolution DNS directe de ce nom et confirmez qu'elle renvoie la même adresse. Voici l'exemple 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
```

Pour traiter de nombreuses adresses à la fois, comparez chacune aux plages IP publiées par Google dans des fichiers JSON en notation CIDR, un fichier par catégorie de client. Les deux vérifications reposent sur le réseau dont provient la requête, que l'expéditeur ne peut pas simplement saisir dans un en-tête.

À quoi cela ressemble en pratique

Le serveur de ce site conserve un journal d'accès, et la page de confidentialité explique comment les robots y sont distingués. Chaque ligne garde l'agent utilisateur, « ce qui permet de séparer les personnes des robots des moteurs de recherche », puis une seconde étape suit : « Un robot qui donne un nom est vérifié par rapport aux réseaux publiés par son propriétaire, à partir de l'adresse masquée décrite ci-dessous. » C'est la seconde méthode de Google, adaptée à ce que le journal conserve. La dernière partie de chaque adresse IP est remplacée par des zéros avant l'écriture sur disque ; l'adresse complète nécessaire à une recherche inverse n'est donc jamais stockée.

Les résumés traitent les robots comme un public distinct : « la même durée pour tout ce qui est automatisé » à côté des trente jours conservés pour les personnes ; un décompte quotidien de « tant de pages lues, tant par des robots, tant qui atteignent la page de contact » ; et un classement mensuel des « robots

et scanners automatisés qui ont accédé au site ». Un robot qui donne honnêtement son nom est compté sous celui-ci ; un nom célèbre envoyé depuis un réseau que son propriétaire ne publie pas échoue à la vérification.

La page sur la préparation aux agents IA prend pour point de départ le même lecteur : « Un agent commence là où commence un robot d'exploration », au fichier robots.txt, au sitemap et aux index de recherche. Elle y répond en faisant de chacun de ces éléments un choix explicite, jusqu'à une annonce IndexNow à chaque déploiement, pour que les moteurs participants « apprennent l'existence d'une nouvelle page le jour de sa publication plutôt qu'à leur prochaine visite ».

Les dates sont l'un des moyens de parler à un robot, mais peuvent finir par ne rien lui dire d'utile. Le récit d'un sitemap qui attribuait le même horodatage à 172 de ses 176 pages expose clairement l'enjeu : « une date de modification promet à un robot que le contenu d'une page a changé ». Un site qui fait cette promesse pour toutes les pages à la fois, poursuit-il, « dit soit la vérité sur une réécriture complète, soit rien d'utile à personne ».

Où les problèmes surviennent

Les requêtes qui se disent Googlebot ne viennent pas de Google. Le journal montre un trafic important sous le nom de Googlebot, et le propriétaire le laisse passer comme un invité bienvenu ou bloque carrément ce nom. Comme d'autres robots le copient souvent, aucune réaction n'est sûre : la première fait confiance à un imposteur, la seconde exclut aussi le vrai robot. Vérifiez le réseau source, par DNS inverse et direct pour une adresse ou par les plages publiées par Google pour un grand nombre, puis décidez sur cette base plutôt que sur le nom.

Une page publiée n'apparaît jamais dans une exploration. Rien ne pointe vers elle, ou elle n'est accessible qu'après connexion. Sans registre central, un robot découvre une adresse par les liens et les sitemaps ; le guide Google cite les pages derrière une connexion parmi celles que son robot ne peut atteindre. Ajoutez un lien depuis une page déjà explorée, inscrivez-la dans le sitemap et laissez accessibles sans connexion les contenus destinés à la recherche.

Le rythme d'exploration baisse après une mauvaise période pour le serveur. Moins de pages sont récupérées chaque jour et les changements mettent plus longtemps à être vus. Les réponses lentes, erreurs 5xx et réponses 429 abaissent toutes la limite de capacité d'exploration de Google ; elle ne remonte que lorsque les réponses redeviennent stables et rapides, et tous les robots Google de l'hôte

partagent ce qu'il en reste. Corrigez d'abord les erreurs et la latence ; pour quelques milliers de pages, le guide du budget d'exploration de Google n'est pas le point de départ.

Une page sans lien apparaît dans les requêtes d'un robot. L'adresse n'a jamais été publiée, peut-être un brouillon privé ou une copie de test, mais figure tout de même dans le journal sous le nom d'un robot. Google juge « presque impossible de garder un site secret en ne publiant pas de liens vers lui », car un clic de cette page vers un autre site peut laisser son adresse dans un journal de provenance qui pourrait être publié. Placez tout contenu privé derrière une protection par mot de passe, le moyen que Google indique pour tenir à l'écart robots et personnes.

Pour aller plus loin

- RFC 9309 : protocole d'exclusion des robots — la conception des robots dans la norme : jetons de produit, chaîne User-Agent et règles qu'on leur demande de respecter.
- Guide détaillé du fonctionnement de la recherche Google — découverte, exploration, rendu et indexation tels que Google les décrit.
- Vue d'ensemble des robots et outils de récupération Google (agents utilisateurs) — les trois catégories de clients Google, les limites de taille, les protocoles et le cache.
- Vérifier les requêtes des robots et outils de récupération Google — recherches DNS et plages IP publiées, avec des exemples détaillés.
- Optimiser le budget d'exploration — comment Google fixe son rythme d'exploration et quels sites doivent le régler.
- Googlebot — robots pour smartphone et ordinateur, fréquence des visites, limite de taille et raison pour laquelle une page sans lien n'est pas secrète.

Glossaire

Choisissez selon vos contraintes réelles

Présentez le processus de travail, les interfaces existantes et les utilisateurs qui ont besoin du changement. Nous pouvons définir une livraison testable et une transmission permettant à votre équipe d'assurer la maintenance des éléments livrés.

Consulter les tarifs Contactez-nous →

Rendez la terminologie concrète

Demandez comment « Qu'est-ce qu'un robot d'exploration ? » s'applique à votre organisation. Un exemple concret tiré de votre système actuel suffit pour engager la discussion.

Découvrir nos prestations

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

Emportez cette page avec vous

Fiche de contact

Créé avec soin et amour par D.

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

Droit d'auteur © 2025 – aujourd'hui · Engineer ApS · Tous droits réservés.

Imprimé depuis https://engineer.company/fr/glossary/crawler/