Engineer

クローラー

🕷 ウェブのリンクを自動でたどり、ページを取得して一覧を広げるプログラム

どのようなものか

ウェブの全ページをまとめた一覧はありません。そのため、検索エンジンは自らページを探す必要があり、そこで送り出すプログラムがクローラーです。既に知っているアドレスから始め、それぞれをHTTPで取得し、返された内容のリンクを読んで、新しいアドレスを一覧に加えます。Googleの検索の仕組みに関するガイドは、別名として「クローラー、ロボット、ボット、スパイダー」も挙げています。ロボット排除の標準であるRFC 9309は、クローラーを、検索エンジンが「インデックス作成のためにリンクを再帰的にたどる」自動クライアントと説明しています。

主な入り口はリンクです。Googleによると、Googlebotは新しいアドレスを「主として、以前クロールしたページに埋め込まれたリンクから」見つけます。ガイドには、新しい投稿へリンクするカテゴリーページという、おなじみの例が載っています。第二の入り口は、サイトの所有者が直接渡す一覧であるサイトマップです。どちらにも載っていないページは、クローラーが存在を知る手掛かりを与えません。

取得は、Googleが説明する三段階の最初です。クロールではページのテキスト、画像、動画をダウンロードします。インデックス作成では、ダウンロードしたものを解析して保存します。検索結果の提供では、そのインデックスから検索に答えます。Googleはクロール中にページの描画も行い、新しいバージョンのChromeでJavaScriptを実行します。それをしなければ、「そのコンテンツが見えない可能性がある」ためです。どの段階も保証されていません。Googleは「ページのクロール、インデックス作成、検索結果への表示を保証しない」とし、「サイトをより頻繁にクロールする対価として支払いを受け付けない」とも述べています。

クローラーには、身元を名乗ることも求められます。RFC 9309では、クローラーごとに自分の名前となるプロダクトトークンを設定します。使えるのは、英字、アンダースコア、ハイフンだけです。HTTPではUser-Agentヘッダーの値となる識別文字列に、このトークンを含め、その文字列でクローラーの目的を説明すべきだとしています。RFCの例では、ヘッダーに `ExampleBot/0.1` とクローラーを説明するページへのリンクがあり、対応するrobots.txtの行は `user-agent: ExampleBot` です。ただし、この名前は自己申告であり、身元の証明ではありません。どのプログラムでも送れるためです。

Googleからの自動リクエストが、すべてクロールとは限りません。Googleの概要は、クライアントを三種類に分けています。Googlebotなどの一般的なクローラーは、「自動クロールでは常にrobots.txtの規則に従う」と説明されています。特殊なクローラーは、サイトとの合意に基づいて動き、その規則に従う場合も従わない場合もあります。利用者の操作で起動する取得プログラムは、通常、人が求めたために一回だけリクエストし、robots.txtは無視します。

概念を判断に生かす

「クローラーとは？」を実際の判断に生かしましょう。その概念を、開発または購入するシステムに結び付けて考えるお手伝いをします。 お客様のAIアシスタントがサイトやアプリを見つけ、読み取り、操作できる状態を、最初の約束だけで済ませず、リリースごとに検証します。

サービスを見る お問い合わせ →

用語の背後にある前提を検証する

解説とリンク先の実績を参考に、プロジェクトについての質問を整理してください。ツールを選ぶ前に、前提を一緒に検討できます。 3言語の多言語Hugoサイトを構築 — 108 3言語・4形式を提供するテンプレートファイル

3言語の多言語Hugoサイトを構築 お問い合わせ →

なぜ重要なのか

再訪する時期はクローラー自身が決め、その振る舞いが、検索で見つけてもらうために負担するコストの一部を左右します。Googleは、大半のサイトでGooglebotが「平均すると数秒に一回より頻繁にはアクセスしないはずだ」としています。サイトに余裕がなくなると、クローラーは頻度を下げます。検索ガイドの言葉では、HTTP 500エラーは「速度を落とす」合図です。クロールバジェットのガイドには仕組み全体の説明がありますが、対象は、週ごとに更新する約百万ページのサイトや、毎日更新する一万ページのサイトです。公開したその日にページがクロールされるサイトには、このガイドは不要だとGoogleは述べています。

クローラーが各ファイルを読む量にも限りがあります。検索用のGooglebotは、対応する種類のファイルでは先頭の2 MB、PDFでは先頭の64 MBを、圧縮前のサイズで読み、それ以降は無視します。

弱点は名前です。Googleは、Googlebotが送るUser-Agentヘッダーが「ほかのクローラーによって頻繁に偽装される」と注意しています。そのため、ログの一行にGooglebotとあっても、それだけでは何も証明しません。Googleは二つの確認方法を示しています。一つのアドレスを調べる場合は、DNSの逆引きを行い、ホスト名が `googlebot.com` 、 `google.com` 、 `googleusercontent.com` のいずれかで終わることを確かめます。その名前を正引きし、同じアドレスが返ることも確認します。Google自身の例は次のとおりです。

```
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
```

多数のアドレスをまとめて調べる場合は、それぞれを、Googleがクライアントの種類ごとにJSONファイルで公開する、CIDR表記のIPアドレス範囲と照合します。どちらの確認も、リクエストの送信元ネットワークを根拠にします。送信者がヘッダーに好きな値を書くだけでは変えられない情報です。

実際の活用例

このサイトのサーバーはアクセスログを保持し、プライバシーページで、その中のクローラーをどう区別するか説明しています。各行にユーザーエージェントを残し、「それによって人と検索エンジンのクローラーを区別します」と説明しています。さらに、第二の手順があります。「名前を名乗るクローラーは、後述の、末尾をぼかしたアドレスを使い、所有者が公開するネットワークと照合します」としています。これはGoogleの二つ目の方法であり、このログが残す情報に適しています。各IPアドレスの末尾は、行がディスクに届く前にゼロへ置き換えます。そのため、逆引きに必要な完全なアドレスは、保存していません。

集計では、クローラーを独立した読者層として扱います。人については三十日間保持する記録の説明に、「自動アクセスもすべて同じ」と添えています。毎日の集計は、「読まれたページの数、そのうちクローラーによるものの数、問い合わせページに到達した数」です。毎月の順位表には、「アクセスしたクローラーと自動スキャナー」を載せます。正直に名乗るクローラーは、自分の名前で数えます。有名な名前を使っていても、所有者が公開していないネットワークから来れば、確認を通過しません。

AIエージェントへの対応に関するページも、同じ読者から考え始めています。「エージェントの出発点は、クローラーと同じ」です。robots.txtファイル、サイトマップ、検索インデックスから始めます。その答えとして、デプロイごとのIndexNow通知に至るまで、それぞれを意図して整備します。参加する検索エンジンが、「次の巡回時ではなく、公開したその日に新しいページを知る」ためです。

日付も、サイトからクローラーへの伝達手段です。ただし、役立つ情報を何も伝えなくなる場合があります。176ページ中172ページに同じタイムスタンプを付けたサイトマップの記録は、問題を明確に述べています。「更新日は、ページの内容が変わったという、クローラーへの約束」です。すべてのページについて一度にその約束をするサイトは、「全体を書き直した事実を伝えているか、誰にも役立つ情報を何も伝えていないかのどちらか」だと続けています。

問題が起きるところ

Googlebotを名乗るリクエストが、Googleから来ていません。 ログにGooglebot名義の大量のアクセスがあり、所有者は歓迎する相手として通すか、名前ごと遮断します。ほかのクローラーがその名前をよく使うため、どちらの反応も安全ではありません。前者は偽物を信用し、後者は本物のクローラーまで締め出します。一つのアドレスならDNSの逆引きと正引き、多数ならGoogleの公開範囲との照合によって、送信元ネットワークを確認します。名前ではなく、その結果で判断します。

公開したページが、一度もクロールされません。 どこからもリンクされていないか、ログイン後にしか到達できません。中央の登録簿はないため、クローラーはリンクとサイトマップからアドレスを知ります。Googleのガイドも、ログインが必要なページを、クローラーが到達できないものとして挙げています。既にクロールされた場所からページへリンクし、サイトマップに載せ、検索に出したいものにはログインを求めないようにします。

サーバーの不調が続いた後、クロールの頻度が下がります。 一日に取得されるページが減り、変更が認識されるまでの時間も延びます。応答の遅さ、5xxエラー、429応答は、いずれもGoogleのクロール処理能力の上限を下げます。応答が安定し、速くなって初めて回復し、残った枠は、そのホストを巡回するすべてのGoogleクローラーで共有します。まず、エラーと遅延を直します。数千ページのサイトなら、クロールバジェットのガイドを出発点にすべきではありません。

リンクしていないページが、クローラーのリクエストに現れます。 アドレスは公開していません。非公開の下書きや、テスト用のコピーかもしれません。それでも、クローラーの名前でアクセスログに現れます。Googleは、リンクを公開しないことでサイトを秘密に保つのは「ほぼ不可能」だとしています。そのページから別のサイトへのクリックにより、参照元ログにアドレスが残り、そのログが公開される場合があるためです。非公開のものは、パスワードで保護します。Googleも、クローラーと人の両方を締め出す方法として、これを挙げています。

参考資料

- RFC 9309：ロボット排除プロトコル — プロダクトトークン、User-Agent文字列、クローラーに従うよう求める規則について、標準がどう定めているかを説明します。
- Google検索の仕組みの詳細ガイド — 検出、クロール、描画、インデックス作成についての、Googleによる説明です。
- Googleのクローラーと取得プログラム（ユーザーエージェント）の概要 — Googleクライアントの三種類、ファイルサイズの上限、プロトコル、キャッシュを説明します。
- Googleのクローラーと取得プログラムからのリクエストの確認 — DNSの照会と公開IPアドレス範囲を、実例とともに説明します。
- クロールバジェットの最適化 — Googleが巡回頻度をどう決め、どのようなサイトで調整が必要かを説明します。
- Googlebot — スマートフォン用とデスクトップ用のクローラー、Googlebotの訪問頻度、読み込むファイルサイズの上限、リンクされていないページも秘密ではない理由を説明します。

用語集

実際の制約に合わせて選ぶ

業務の流れ、既存のインターフェース、変更を必要とする利用者をお知らせください。検証可能な納品内容と、お客様のチームが保守を続けられる引き継ぎ内容を定められます。

料金を見る お問い合わせ →

用語を実務に結び付ける

「クローラーとは？」がお客様の組織にどう当てはまるかをご相談ください。現在のシステムの具体例が一つあれば、話を始められます。

サービスを見る

welcome@engineer.company +45 50 10 61 84 デンマーク、コペンハーゲン

このページを持ち歩く

連絡先カード

心を込めて丁寧に制作： D.

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

著作権 © 2025 – 現在 · Engineer ApS · すべての権利を留保します。

印刷元： https://engineer.company/ja/glossary/crawler/