# robots.txtを効果的に設定しよう

## クロールが不要な部分はrobots.txtで回避する

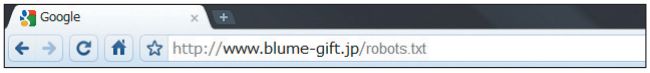
**“robots.txt”とは、検索エンジンにアクセスしクロールしてほしい部分と、そうでない部分を伝えるためのファイルです**（❶）。このファイルは必ず“robots.txt”というファイル名でサイトのルートディレクトリに置く必要があります（❷）。

Googleウェブマスターツールをご利用いただくと、robots.txtファイルを簡単に作ることができます。詳細は、ヘルプセンターのrobots.txtファイルを使用してページをブロックまたは削除するをご確認ください。サブドメインを持つサイトで、ある特定のサブドメイン内のページをクロールさせないようにするには、そのサブドメイン用に別のrobots.txtファイルを用意する必要があります。

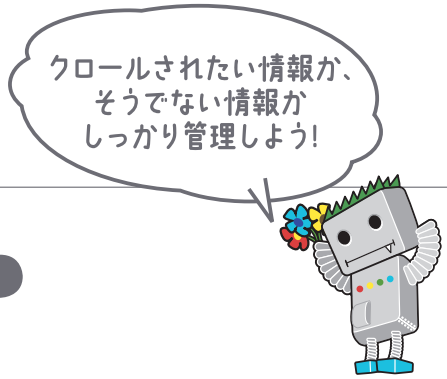
検索結果にコンテンツを表示させない方法は他にも、“NOINDEX”をrobotsメタタグに追加、.htaccessを使ってディレクトリにパスワードを設定、Googleウェブマスターツールを使ってすでにクロールされたコンテンツを削除するなどがあります。

```
User-agent: *
Disallow: /image/
Disallow: /search
```

❶Robots Exclusion Standardに準拠している検索エンジンのロボットすべて（「*」というワイルドカードのシンボルで表現される）に対し、/image/以下にあるコンテンツ、もしくは/searchで始まるURLにあるコンテンツに、アクセスもクロールもさせない場合の例




❷フラワーギフトショップのrobots.txtファイルのアドレス




### ポイント

## 慎重に扱うべきコンテンツにはより安全な方法を使用しよう

機密事項や慎重に扱うべきコンテンツがクロールされないようにするには、robots.txtの設置だけでは十分ではありません。その理由の1つは、クロールできないように設定したURLであっても、そのURLへのリンクがインターネット上のどこか（例えばリファラーログなど）に存在する場合、検索エンジンはそのURLを参照できるからです。また、Robots Exclusion Standardに準拠しない検索エンジンや不正な検索エンジンなどは、robots.txtの指示に従わないかもしれません。そしてもう1つ、好奇心の強いユーザーの中には、robots.txtにあるディレクトリやサブディレクトリを見て、見られたくないコンテンツのURLを推測する人がいるかもしれません。コンテンツの暗号化や.htaccessを使ってパスワードをかけて保護する方が、より確実で安全な手段だといえます。

- ☑ **検索結果のようなページはクロールさせない**
  ※検索結果のページから、さほど価値が変わらない別の検索結果のページへ飛んでも、ユーザーの利便性を損なうだけです
- ☑ **同一か、ほとんど違いがない自動生成されたページを大量にクロールさせないようにする**
  ※「重複コンテンツに近いこれら100,000ものページはインデックスされるべきだろうか？」と問い直してみましょう
- ☑ **プロキシサービスによって生成されたURLはクロールさせないようにする**

**.htaccess**
ウェブサーバーの動作環境を制御するアクセス環境設定ファイル

**リファラーログ**
アクセスログに記載されているリファラー情報。これをたどっていくと閲覧者がどこのサイトから来たかなどを調べられる

**プロキシサービス**
内部ネットワークと外部ネットワークを接続する場合に接続を代行するコンピュータ、またはそのための機能を持ったソフトウェアのこと

**参考ページ**

ウェブマスター向けヘルプセンター
http://www.google.co.jp/support/webmasters/
↳ robots.txtファイルを使用してページをブロックまたは削除する 検索