



Google 検索アプライアンス - フィードプロトコル 機能の概要

ビジネスの概要

新しい Google 検索アプライアンスのフィードプロトコルでは、シンプルな XML 変換により、ウェブ上でアクセスできないコンテンツを Google 検索アプライアンスに送り込むことができます。この機能により、企業内の情報利用の障壁を取り除き、あらゆる企業コンテンツに簡単にアクセスして、検索することができます。フィードプロトコルにより、企業管理者はドキュメント管理、企業内アプリケーション、または旧システムなどのソースにある、ウェブ上でアクセスできないコンテンツを直接 Google 検索アプライアンスに送り込み、インデックス登録および検索を実行できるようになります。

コンテンツのエクスポート

企業コンテンツを Google 検索アプライアンスに送り込むには、まず最初にレコードシステムから企業コンテンツをエクスポートし、Google 検索アプライアンスへの転送に適したフォーマットに変換する必要があります。Google 検索アプライアンスのフィードプロトコルでは、業界標準の XML (Extensible Markup Language) を使用して、シンプルで簡潔なフィードを生成します。

Google 検索アプライアンスのフィードプロトコルでは、コンテンツ フィードと URL のみのフィードの 2 種類のデータ フィードがサポートされます。各ドキュメントは固有のキーとしてその URL により追跡され、追加のパラメータは XML フィード ファイルで属性として設定されます。

コンテンツ フィード

コンテンツ フィードでは、ドキュメントのコンテンツを元の形式 (HTML、テキスト、Microsoft Word、PDF、Microsoft Powerpoint など) で Google 検索アプライアンスに送り込むことができます。これにより、ウェブ上でアクセスできないドキュメントを検索アプライアンスのインデックスに登録できます。HTML やテキスト ドキュメントなどのテキスト ベースのコンテンツは、XML フィード ファイルの本文の <content> タグに直接挿入されます。Word、Excel、PDF ファイルなどのバイナリドキュメントは、base64 エンコーディングされ、エンコード後のテキストが XML フィード ファイルの本文の <content> タグに挿入されます。ファイル形式は、<record> 要素の mimetype 属性で指定されます。

URL のみのフィード

URL のみのフィードでは、HTTP でアクセス可能な URL のリストを Google 検索アプライアンスのウェブ クローラに送信して、クローラおよびインデックス登録を行うようにすることができます。これにより、自動的に見つけるのが難しく、通常のウェブ

仕様

関連技術:

XML – Extensible
Markup Language

HTTP – ハイパーテキスト
転送プロトコル

Google 検索アプライアンス

ハードウェア: バージョン 4.0 以降
ソフトウェア: バージョン 4.2 以降

お問い合わせ

www.google.co.jp/enterprise

もしくは同ページ [お問い合わせ] よりお問い合わせください

Google 検索アプライアンス

クローラでクロール可能な URL を検索アプライアンスに提供できます。URL はウェブ クローラのキューの先頭に追加され、ウェブ クローラの稼動時にページがクロールされ、インデックスに登録されます。

コンテンツの取り込み

必要なコンテンツ フィードの生成が完了したら、HTTP-POST メソッドを使用して、Google 検索アプライアンスに XML フィード ファイルを送り込みます。固有のコンテンツ レポジトリはデータ ソースとして参照され、レポート用にわかりやすい名前を付けることができます。フィード プロトコルでは、完全フィードと段階的フィードの両方がサポートされます。POST メソッドの呼び出しは、データ ソース名、フィード タイプ、XML ファイルのパスで構成されます。

例:

```
http://<search_appliance>:19900/xmlfeed/datasource=&feedtype=&data=<xmluri>
```

完全フィード

完全フィードでは、完全なデータ ソースを送り込みます。つまり、アプライアンスに完全フィードを送り込むと、データ ソースの全ドキュメントが送られることとなります。アプライアンスでは、完全フィードに含まれなくなったドキュメントを特定し、インデックスから削除します。

例:

顧客が完全フィードを作成し、アプライアンスに送り込んだとします。

- 完全フィードには、ドキュメント D0、D1、D2 が含まれています。システムのインデックスには、D0、D1、D2 と他のクロールしたドキュメントが登録されています。
- その後、管理者がドキュメント D0、更新された D1、新しい D3 を含む別の完全フィードを作成し、アプライアンスに送り込んだとします。
- フィードの処理が完了すると、システムのインデックスには、D0、更新された D1、新しい D3 と他のクロールしたドキュメントが登録されます。
- ドキュメント D2 は 2 つ目の完全フィードに含まれていないため、インデックスから削除されます。

段階的フィード

段階的フィードでは、フィードでドキュメントを削除して、新しいドキュメントを送り込みます。このため、システムにとってより効率的で、大規模なデータ ソースの制約にも対応することができます。一般的な使用方法としては、最初は完全フィードを送り込み、その後の更新では段階的フィードを送り込みます。完全フィードは、フィードのコンテンツを "リセット" するためにいつでも使用できます。

例:

- 1 日目 - 管理者がドキュメント D0、D1、D2 を含む完全フィードを送り込みます。システムのインデックスには、D0、D1、D2 が登録されます。
- 2 日目 - 管理者がドキュメント D3 の "追加"、更新した D1 の "追加"、D2 の "削除" を含む段階的フィードを送り込みます。システムのインデックスには、D0、更新さ

Google 検索アプライアンス

れた D1、D3 が登録されます。

- 7 日目 - 管理者がドキュメント D0、D7、D10 を含む完全フィードを送り込みます。完全フィードの処理が完了すると、システムのインデックスには、D0、D7、D10 が登録されます。

結果の配信

Google 検索アプライアンスのフィード プロトコルを介して受け取ったドキュメントは、検索アプライアンスのインデックスに登録されます。ユーザーが検索を行うと、フィード送信したドキュメントのほか、データベースのコンテンツやクロールしたコンテンツも結果に表示されます。フィード送信したドキュメントにウェブアクセスが可能な URL が設定されている場合、ユーザーはその URL を選択して元のドキュメントにアクセスできます。フィード送信したドキュメントがウェブ上またはブラウザでアクセスできない場合は、キャッシュされたドキュメントを表示したり、ドキュメントの URL を使用して他の方法でソース ドキュメントにアクセスできます。