Keywalker Web Crawler

ご説明資料

株式会社キーウォーカー

2012年５月８日

Webクローラの利用シーン例

- マーケティング、価格調査時の情報収集
  - 口コミ評判調査/価格調査/競合調査/など
- ポータル生成時の情報収集
  - グループサイト全てを統合したポータルの作成
  - 社内Webなどのポータル作成
  - キーワードまとめサイトの構築
  - サイト内検索システム構築
- Webの情報収集
  - 営業情報収集/コンテンツ用情報収集など

良いクローラはないか？

- アルバイトを使って競合先のWebサイトを調査
- 検索エンジンで市場調査

大変/コスト高/更新に追いつけない

自動化

- フリーソフトやWget コマンドで、Webページを収集しデータ加工

取得後のデータ加工に手間暇が掛かる

(・ ・;)

後処理が大変で機能もイマイチ！！

開発すると数百万円かかって、納期もかかる、インフラも自前持ちで運用も大変! ➡ ならアルバイトで調査したほうが安いかも…

Keywalker Web Crawler

他では実現できない高機能汎用Webクローラを
お届けします！

Keywalker

12/05/08

Keywalker Web クローラとは

- Web情報の収集ツール
- 高機能な収集情報分類機能
- ASPなので、安価ですぐ収集開始
- クロールスケジュールなどを細かく設定でき、更新情報を逃しません。
- サーバの性能や台数を調整できるので、大量の情報でも効率よく収集できます。

一般クローラとの比較

Wget や 一般のSpider Robot

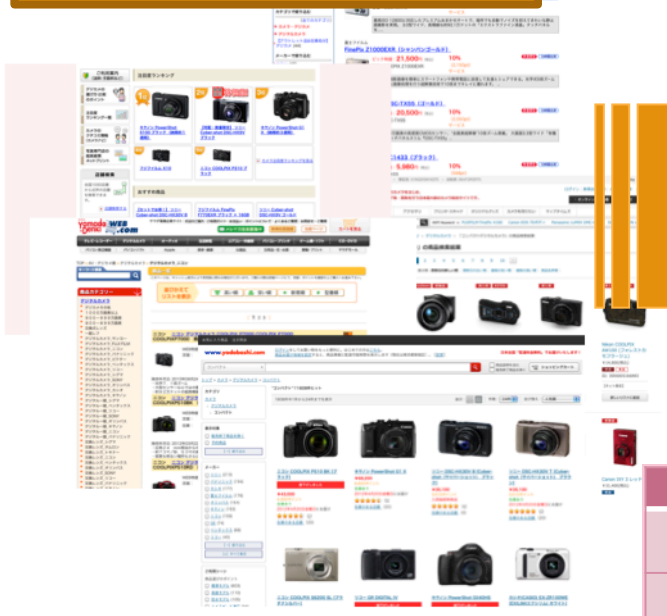
- 収集情報はHTMLファイル保存
- 必要な情報は、保存ファイルからの抽出が必要
- 不要なリンクまで全てクロール
- Webページを収集するのみ

Keywalker Web Crawler

- 収集情報は、TAGを自動除去しテキスト保存
- 蓄積情報を分類データとしてCSVでダウンロード
- 高効率なクローリングルート設定
- 収集情報をキーワード検索したり、Webで確認することが可能
- オプションで特徴語抽出など日本語処理技術を提供

市場価格調査の例

対象コマースサイト群



クロール

KW Web クローラ

取得サイトの設定情報

分類蓄積

CSV

サイト	取得日	商品名	価格	コメント
淀カメラ.com	2012/04/25 10:30	DIGITAL IV	¥61,300	値下げしました
大カメラ.jp	2012/04/25 10:30	DIGITAL IV	¥61,800	特価
山電機.net	2012/04/25 10:30	DIGITAL IV	¥64,800	特価
地図カメラ	2012/04/25 10:30	DIGITAL IV	¥72,500	アクセサリ付特化
土居	2012/04/25 10:30	DIGITAL IV	¥72,800	Web特価

Keywalker Web Crawler の主な機能

- HTML以外の多種類のデータを取り扱い可能
- 効率の良いクロールルートの設定が可能
- きめ細やかな間隔やスケジュールの調整が可能
- 細かなデータの分類取得設定が可能

取得データの種類

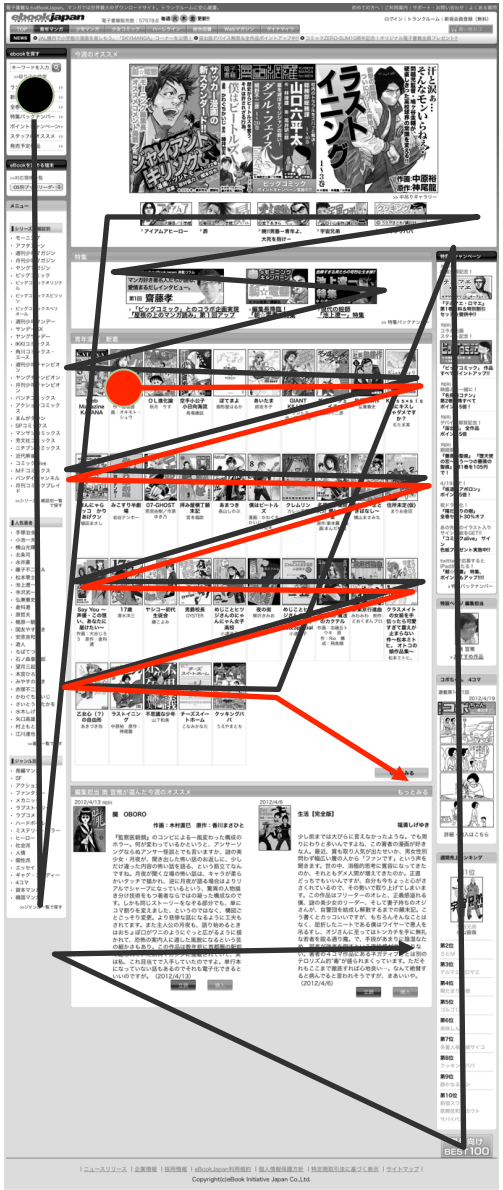
- KW Web クローラは、HTMLだけでなく多くのファイルフォーマットをサポートします。

フォーマット	内　　容
HTML	Webページ記述フォーマット
RSS	RSS更新情報➡RSSから本体サイトの情報をクロールします。
SITEMAP	Webサイトのサイトマップから本体情報をクロールします。
PDF	PDFのテキスト部分を抽出します。
DOC	Microsoft Word のテキスト部分を抽出します。
XLS	Microsoft Excel のテキスト部分を抽出します。
PPT	Microsoft Power Point のテキスト部分を抽出します。

クロールルートの調整

- 通常のクローラは、URLとクロール深度設定のみなので、不要なリンクをたどって無駄な情報を取得します。
- KW Web クローラは、必要なルートのみたどれるように設定できます。

一般クローラ	KW Webクローラ
全てのリンクをたどるので無駄が多く、クロール対象サーバへの負荷も高い。	必要なリンクしかたどらないため、処理が早く、対象サーバへ無駄な負荷を掛けない。



右の図はクロールルートのイメージ
クローラはルート上のすべてのリンクをたどって必要な情報へ遷移する。
一般のクローラでは、無駄なページを取得するため、時間/ハードリソース/ネットワークリソースを大量に消費する。
KW Webクローラは、必要な部分しかたどらないらないため、効率が高い。

Keywalker

クロール間隔/スケジュールの調整

- きめ細やかなスケジューリング機能で、データの更新タイミングを逃しません。

その他の機能

- 確認画面レイアウト機能
  - 確認画面をレイアウトする機能
- データ確認画面
  - Webページで取得データを検索し、確認できます。
- ファイルダウンロード機能
  - 取得したデータをCSVダウンロード
- 特徴語抽出機能（オプション）
  - 取得したデータからページ毎に特徴語を抽出する機能
- 連携API（オプション）
  - データ利用サービスに合わせたAPI作成

実　績

- ４０社以上の導入実績
- 検索サービスとしての実績
  - AFP通信社 http://afpbb.com/ http://afpbb.com/fashion/
  - 日刊工業新聞社 http://www.nikkan.co.jp/
  - ウォール・ストリート・ジャーナル http://jp.wsj.com/
- キーワードページまとめページとしての実績
  - 毎日新聞デジタル http://keyword.mantan-web.jp/
  - ネクスト ロココム http://lococom.keywalker.jp/

- 上記のサイトには、KW Web クローラが使われています。

Keywalker Web Crawler

お問い合わせ、ご用命は下記まで御連絡下さい。

〒106-0041
東京都港区麻布台２−４−２

株式会社キーウォーカー
電話：０３−３５６０−６２０１
Mail：sales@keywalker.co.jp
担当：山本

http://www.keywalker.co.jp/