Fessでつくる、ニッチ検索エンジンの可能性

2015/04/25　Y.K. Works

Subject matter (今日の題材)

○ 観音寺市 情報検索エンジン
(http://search.kanonji.si/)

観音寺市に関する情報に絞って検索できるようにした、ニッチ用途向けサーチエンジン。テスト運営中。

市役所からの情報やグルメ情報はもちろん、観音寺市が舞台ということで(一部で)話題になった「結城友奈は勇者である」の聖地情報などもカバー。

What's Fess？(Fessってなに？)

- Javaで作られたオープンソースの全文検索システム(java + tomcat)
- 5分で環境の構築が可能(Javaは別途インストールが必要)

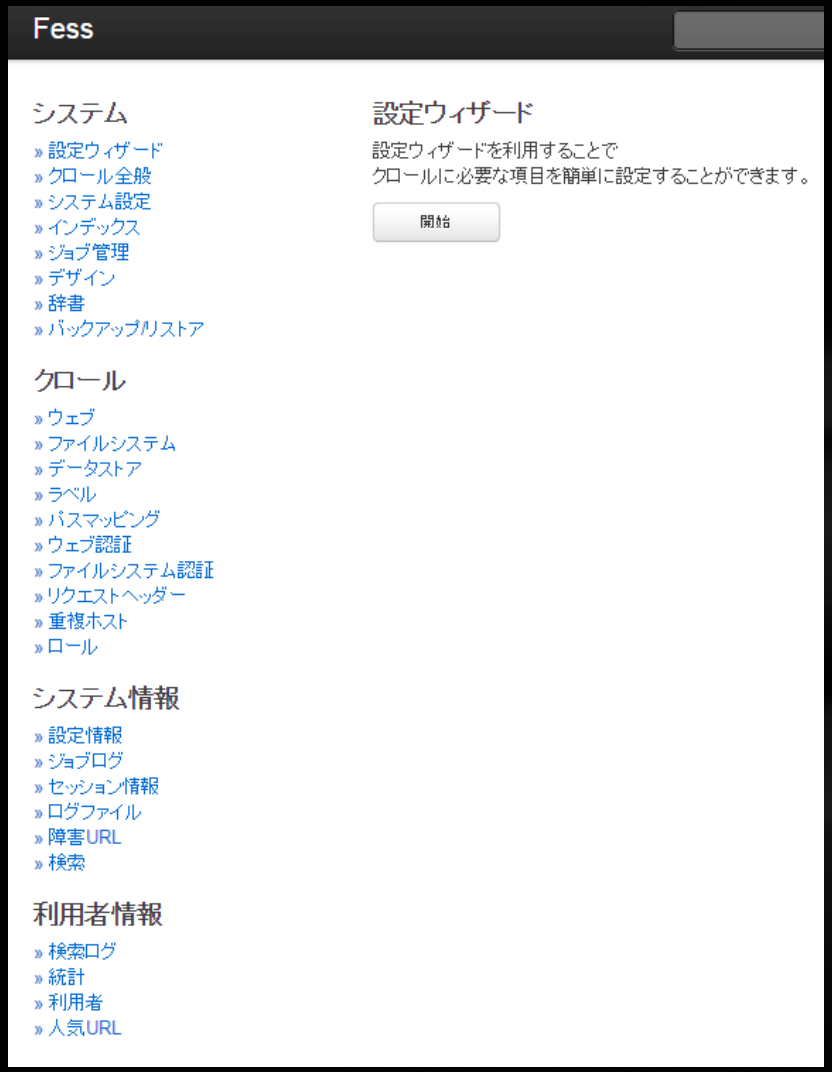
```
$ unzip fess-server-x.y.zip
$ cd fess-server-x.y
$ chmod +x bin/*.sh
```

- Webクローラーも付いている、オールインワンパッケージ

Fess Website (詳細はこちらで)

http://fess.codelibs.org/ja

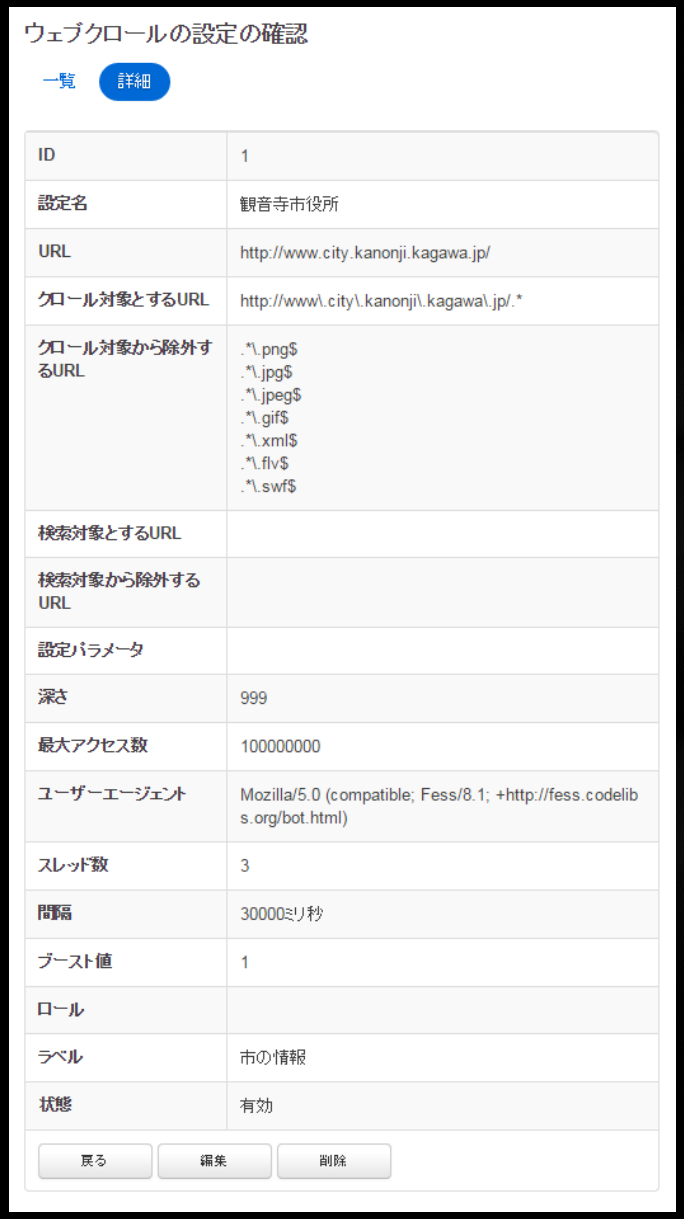
Settings (設定画面)



Point

- ウィザードに従うと、とりあえず使い始めることができる
- 詳細なクロール設定は、
クロール – ウェブ
から設定する

Example 1. (クロール設定例 1)


ウェブクロールの設定の確認

一覧 詳細

ID	1
設定名	観音寺市役所
URL	http://www.city.kanonji.kagawa.jp/
クロール対象とするURL	http://www\.city\.kanonji\.kagawa\.jp/.*
クロール対象から除外するURL	.*\.png$.*\.jpg$.*\.jpeg$.*\.gif$.*\.xml$.*\.flv$.*\.swf$
検索対象とするURL	
検索対象から除外するURL	
設定パラメータ	
深さ	999
最大アクセス数	100000000
ユーザーエージェント	Mozilla/5.0 (compatible; Fess/8.1; +http://fess.codelibs.org/bot.html)
スレッド数	3
間隔	30000ミリ秒
ブースト値	1
ロール	
ラベル	市の情報
状態	有効

戻る 編集 削除

Point

- クロール対象・除外部分は正規表現で指定OK
- クロール対象URLを指定しないと、際限なくクロールするので注意
- ブースト値で検索時の重み付け、ラベルでカテゴリ分類ができる

Example 2. (クロール設定例 2)

ID	21
設定名	結城友奈は勇者であるシングルページ取得
URL	http://anitabi.net/blog/2014/10/yuyuyu.html http://www.nicovideo.jp/watch/sm24684695 http://etesuke.blog.fc2.com/blog-entry-560.html http://tsurebashi.blog123.fc2.com/blog-entry-371.html http://tsurebashi.blog123.fc2.com/blog-entry-376.html http://tsurebashi.blog123.fc2.com/blog-entry-379.html http://tsurebashi.blog123.fc2.com/blog-entry-384.html http://tsurebashi.blog123.fc2.com/blog-entry-385.html http://tsurebashi.blog123.fc2.com/blog-entry-386.html http://tsurebashi.blog123.fc2.com/blog-entry-387.html http://tsurebashi.blog123.fc2.com/blog-entry-388.html
クロール対象とするURL	
クロール対象から除外するURL	.*\.png$.*\.jpg$.*\.jpeg$.*\.gif$.*\.xml$
検索対象とするURL	
検索対象から除外するURL	
設定パラメータ	
深さ	0
最大アクセス数	
ユーザーエージェント	Mozilla/5.0 (compatible; Fess/9.2; +http://fess.codelibs.org/bot.html)
スレッド数	3
間隔	30000ミリ秒
ブースト値	1
ロール	
ラベル	舞台作品
状態	有効

Point

- 深さを0とすることで、指定したURL 1ページのみ取得できる
- 1ページのみ取得させたい場合は、クロール対象とするURLは指定しなくてもOK

Request Header (リクエストヘッダー機能)

- クロール時に、任意のHTTPリクエストヘッダーを相手に渡すことができる機能
- 一般的には、シングルサインオン環境のファイルサーバーのクロールや、社内ポータルサイト等の検索に使う(のかな？)
- 応用すると、ちょっといい感じのこともできます

In other words？(つまり何ができるの？)

- 特定のCookieを要求するサイトもクロールできる

- リクエストヘッダー以外では、BASIC・DIGEST・Active Directory認証にも対応しているので、そういったサイトもクロール可能

Request Header Example 1. (設定例 1)

Point

- 名前にCookieと指定し、値に
  Cookieの名称＝Cookieの値
  とすることで、Cookie情報をサイトに渡すことが可能

Request Header Example 2. (設定例 2)

Point

- 複数のCookieを指定する場合は、*Cookie名称＝Cookie値;Cookie名称＝Cookie値*とセミコロンで繋ぐことで、渡すことが可能

Summary (まとめ)

- FessはJava+Tomcatベースなので、メモリを多めに積んでないと実運用は厳しい(2GB VPSだとたまに落ちる)
- Googleカスタム検索機能でカバーできる事は多いけど、Googleに頼りすぎると、裏切られたときにダメージが大きい
- 管理者がキュレーションするニッチな検索エンジンは、色々と可能性があるかも