J-02　ロボット型検索エンジンの作成に関する基本研究

千葉　翔太
指導教員　ソソラバラム　バドゥジャルガル

1.　テーマ選定理由

検索エンジンの仕組みに関して興味を持ち、先輩[1]の研究論文を参考にして、自分専用の検索エンジンを動作させたいと思い、このテーマを設定しました。

2.　研究概要

本研究の目的は、検索エンジンの仕組みを理解し、専用の検索エンジンをパソコン上で動作させることです。

2.1 ロボット型検索エンジンの仕組み

クローラと呼ばれるロボットプログラムがインターネットを巡回し、ウェブページの情報を収集します（図 1 参照）。

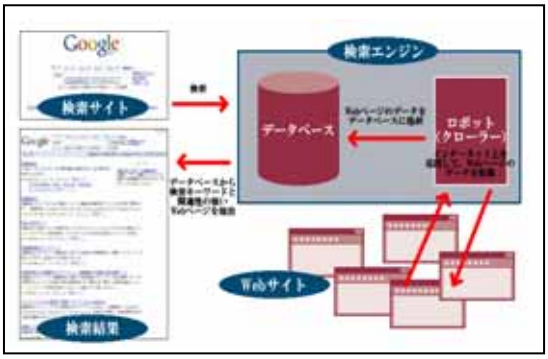

図 1.　ロボット型検索エンジンの仕組み

2.2 My 検索エンジン

Google などのロボット型検索エンジンでは、すべてのウェブページを検索対象にするためデータの量が膨大になります。本研究で動作させる My 検索エンジンでは、検索対象となるお気に入りのサイトをあらかじめ登録しておきます．ロボットプログラムがこの登録されているウェブページの情報を取集します（図 2 参照）。

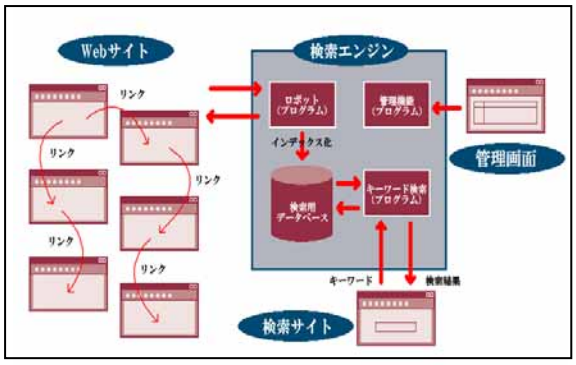

図 2.　My 検索エンジンの仕組み

3.　動作確認

My 検索エンジンの動作の流れを説明します。

ステップ 1　サイトの登録

学内ページ http://www2.iwate-it.ac.jp を、検索対象のページとして登録しました。

ステップ 2　巡回ログ

ロボットが巡回した記録を、図 3 に示す。HTML ファイルの探索の深さを n=7 と設定しました。

ステップ 3　データベースへの蓄積

ロボットが巡回した結果、図 4 で示したとおりに、1186 個のリンクをたどって、合計 80085 個のキーワードが蓄積されました。

- 登録完了ページ
  - Level 1 -- 1
  - Level 2 -- 17
  - Level 3 -- 387
  - Level 4 -- 246
  - Level 5 -- 313
  - Level 6 -- 142
  - Level 7 -- 80
  - Total -- 1186
- キーワード数
  - Total -- 80085

図 4.　システム状況

ステップ 4　キーワードの検索

ユーザー画面からキーワードを入力して検索を行います。その結果として指定したキーワードを含むページ(図 5 参照)がリストアップして表示されます。

4. まとめ

[1], [2]に基づき、小規模な検索エンジンについて学習し、検索エンジンを自分のパソコン上でインストールし、動作の確認を行いました。

ロボットプログラムの流れの確認のため、Ruby 言語[4], 正規表現[3]について学習しました。

今後の課題として、ロボットプログラムのさらなる理解、検索エンジンの用途の検討などがあげられます。

参考文献

[1] 與羽　祐也　“ロボット型検索エンジンの作成に関する基本研究”, 2010.

[2] 星澤　隆　“Ruby で作る検索エンジン”, 2009.

[3] 杉山　貴章　“正規表現書き方ドリル”, 2011.

[4] (株)アンク　“Ruby の絵本”, 2008.

```
2012-02-07 16:31:17: http://www.iwate-it.ac.jp/blog/my_first_blog/assets_c/2010/02/自動車-61.html [1069] parsed.
2012-02-07 16:31:17: http://www.iwate-it.ac.jp/blog/my_first_blog/assets_c/2010/02/etロボコン-56.html [1070] parsed.
2012-02-07 16:31:17: http://www.iwate-it.ac.jp/blog/my_first_blog/assets_c/2010/02/wiiリモコン-57.html [1071] parsed.
2012-02-07 16:31:37: Error Robot.get(5) http://pixta.jp/robots.txt ../lib/suzaku_lib.rb:283:in `get'robot.rb:785:in `get'robot.rb:900:in
`check'robot.rb:1171robot.rb:1116:in `each'robot.rb:1116robot.rb:1070:in `each'robot.rb:1070:Net::HTTP.get timeout.
2012-02-07 16:31:57: Error main(2) http://pixta.jp/@favorreef/ [1072] ../lib/suzaku_lib.rb:283:in `get'robot.rb:1180robot.rb:1116:in
`each'robot.rb:1116robot.rb:1070:in `each'robot.rb:1070:Net::HTTP.get timeout.
2012-02-07 16:31:57: time over.
```

図 3. 巡回ログの一部抜粋

図 5. 検索結果