

# J-18 ロボット型検索エンジンの作成に関する基本研究

2年29番 與羽 祐也

指導教員 ソソラ

## 1. はじめに

インターネットの普及により、多くの情報をウェブページから得られるようになった結果、必要な情報を探すために用いられる“検索エンジン”的重要性がますます高まっています。そこで、検索エンジンの動作やプログラムがどのようなものなのか具体的に理解し、小規模な検索エンジンを動かしてみたいという理由で本テーマを選定しました。

本研究の目的は検索エンジンの仕組みを理解し、検索の対象となるウェブページを自由に設定できるような自分専用の My 検索エンジンをパソコン上で運行してみることです。

以下、まず検索エンジンの仕組みを説明し、前回の発表での研究計画（参照 3 章）において得た研究成果と今後の課題について述べます。

## 2. 検索エンジンの仕組み

私たちに馴染みの深い検索エンジン「Google」の場合、ロボットやクローラーと呼ばれる検索エンジンプログラムがインターネット上を巡回し、ウェブページの情報を収集します。

そして、その収集したウェブページの中からユーザーが指定したキーワードと関連性の高いページが順に表示されます。

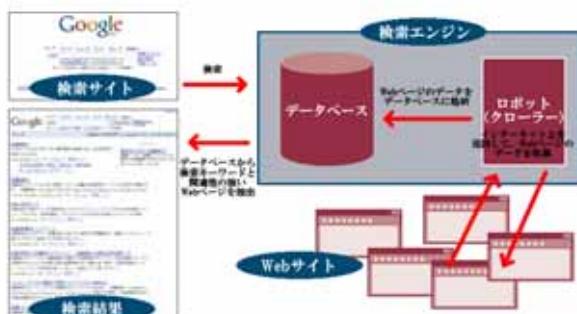


図 1 Google の大まかな仕組み

## 3. 研究計画

前回の中間発表での研究計画は以下の通りでした。

- 開発環境の準備
- 参考文献に基づき小規模な検索エンジンを動かし、動作の確認
- 索引用のテーブルの構造
- オブジェクト指向 Ruby 言語の基礎

### 3.1 開発環境の準備

開発環境の準備として以下の順番で、参考文献 2 の通りインストールを行いました。

- Apache (Web サーバー)
- MySQL
- Ruby
- Ruby + MySQL (MySQL/Ruby for Windows)
- Chasen
- Ruby + Chasen (Ruby/Chasen Alternative)

### 3.2 検索エンジンの動作の確認

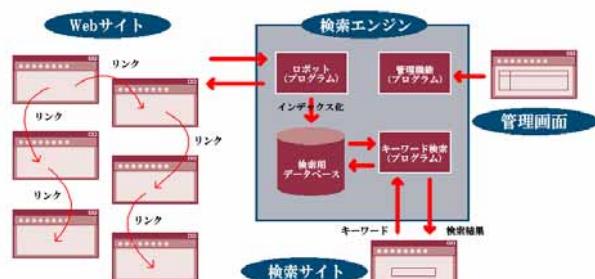


図 2 本検索エンジンの仕組み

小規模な検索エンジンでは一般的な検索エンジンと違い、お気に入りのウェブページを検索対象として設定します。

検索エンジンのプログラムが検索対象のページのみを巡回し、ウェブページの情報を収集します。

収集した HTML ファイルからタグ部分を取り除き、検索対象となる文字列を取り出します。文字列を単語に分解したものと、それを含む HTML ファ

イルの URL をデータベースに保存します。その単語が URL を求めるための索引の役目を果たします。

検索エンジンは 索引となるデータベースを検索して、指定されたキーワードに一致する単語を含む HTML ファイルの URL を特定します。

以下のステップによりエンジンの動作の確認を行いました。

### ステップ 1. データベースやテーブルの作成

	registered	word	count
1	registered	registered	32
2	trademark	trademark	32
3	Inc	Inc	32
4	profit	profit	32
5	us	us	32
6	Privacy	Privacy	32
7	policy	policy	32
8	Disclaimers	Disclaimers	32
9	Powered	Powered	32
10	MediaWiki	MediaWiki	32

### ステップ 2. サイトの登録

### ステップ 3. ロボットの稼動

### ステップ 4. データベースへ蓄積

Level	Count
1	32
2	15
3	12
Total	19431

MySQL Command Line Client				
registered	32	18	2	
trademark	32	18	2	
Inc	32	28	2	
profit	32	18	2	
us	32	26	2	
Privacy	32	26	2	
policy	32	26	2	
Disclaimers	32	26	2	
Powered	32	26	2	
MediaWiki	32	26	2	

### ステップ 5. キーワードを検索

### ステップ 6. 検索の結果を表示

### 4. おわりに

研究計画通りに得られた結果は次の通りです。

- ( 1 ) 開発環境の準備や参考文献に基づき小規模な検索エンジンを動かし、動作の確認を行いました。
- ( 2 ) データベースの構造や各テーブルの中に格納されているデータを確認しました。
- ( 3 ) Ruby 言語の基礎、特に文字列の処理と正規表現に関して学びました。

今後の課題として、ロボット処理の流れ、クラスの設計などの詳細検討があります。

### 参考文献

1. “体系的に学ぶ検索エンジンのしくみ”, 2008
2. “Ruby で作る検索エンジン”, 2009
3. “データベースの基礎”, 2009
4. “Ruby の絵本”, 2008