Rastによる全文検索アプリケーションの構築

株式会社 ネットワーク応用通信研究所 橋本 将

目次

Rast とは

- N-gram 方式の全文検索システム
- 平成16年度 IPA オープンソースソフトウェア活用基盤整備事業採択事業

Rast の特徴

- 検索対象となる文書の分野や言語を選ばない
- 文書情報に付随する属性値の登録
- 多様なファイル形式への対応
- 選択可能な文字列処理モジュールの提供
- 全文検索ライブラリの提供

検索対象の分野や言語を選ばない

- N-gram 方式なので検索漏れが生じない
- N-gram 方式なので辞書の整備が必要ない
- => 検索対象の分野や言語を選ばない

文書情報に付随する属性値の登録

- 属性値は本文と一緒に登録する情報
- 属性値には文字列･整数･日付･日時型がある
- => 検索結果の絞り込みや範囲検索が可能

多様なファイル形式への対応

- 自動で本文や属性値をファイルから取得可能
- 自動で取得するための仕組みがある
- => 新しいファイル形式への対応が容易

選択可能な文字列処理モジュール

- N-gram の切り出し部分をモジュール化
- 分かち書き方式に対応するモジュールを追加
- => 他の方式も実現可能 (後述)

全文検索ライブラリの提供

- C と Ruby で全文検索ライブラリの提供
- Perl や PHP からも利用できるようにして頂きました
- => アプリケーションの作成が容易

Rastの機能

- 基本機能のライブラリ提供
- 外部サーバからの検索・登録
- 複数の検索情報データベースからの検索
- フィルタでの本文、属性データの抽出
- 文字列処理モジュール

基本機能のライブラリ提供

- 検索情報データベース作成機能
- 文書登録機能
- 文書検索機能
- 文書除外機能
- 文書更新機能
- 再構成機能

外部サーバからの検索・登録

- 外部サーバに対して検索・登録が可能
- 基本ライブラリとインターフェイスが同じ

複数の検索情報データベースからの検索

- 複数のデータベースを1つのものとして扱う
- ローカルでも外部サーバでも同じ様に扱える
- スコアを補正してソート可能

フィルタでの本文、属性データの抽出

- 自動で本文や属性値をファイルから取得
- 本文と属性値を解析する必要がない
- フィルタは2種類ある
  - 文書ファイルをテキスト化するフィルタ
  - テキストを処理するフィルタ

テキスト化するフィルタ

- 文書ファイルから直接本文と属性値を抽出
- zip → html などのように多段に使用可能

テキストを処理するフィルタ

- テキスト化するフィルタで抽出した本文を加工する
  - 行末と行頭の日本語を改行文字を除いて登録
    - (例) 全文<改行>検索 => 全文検索
  - 行末と行頭の英単語をハイフンを除いて登録
    - (例) eng-<改行>ine => engine
- 検索洩れを減らす

文字列処理モジュール

- 文字エンコーディング別のトークン分割モジュール
- モジュールの追加で任意の文字エンコーディングに対応可能
  - きちんとした多言語対応など

文字列処理モジュールを追加するには

- 文字エンコーディングを正しく扱う
  - 何バイトで1文字を表現するか
- トークンの切り出し
  - N-gram
  - わかち書き
- 正規化
  - 大文字を小文字に

多言語対応への課題

- 課題
  - 異体字処理など
- 解決策
  - 実装するための仕組みはある
  - 多言語対応な文字列処理モジュール
- 実装はされていないのでよろしくお願いします

各プログラミング言語からの利用

- Ruby
- Perl
- PHP

アプリケーションの実例

- Web 検索 CGI
- morq
- ximapd
- gigi
- tdiary-rast
- pg_rast
- mod_search_rast

Web 検索 CGI

- ブラウザ上で検索結果を表示
- 外部サーバや複数のデータベースが利用可能
- http://projects.netlab.jp/rast/

morq

- 検索ベースの電子メールソフト
- フォルダによる分類を行わず、全部検索する
- ラベル指定が可能
- http://projects.netlab.jp/rast/

ximapd

- 検索ベースの IMAP サーバ
- 検索によって仮想的なメールボックスを提供
- http://projects.netlab.jp/ximapd/

gigi

- MailDir 形式のメールボックスの全文検索ツール
- Perl での利用例としても参考になる
- http://tech.yappo.jp/rast/

tdiary-rast

- Rastで検索可能になるtDiary 用のプラグイン
- 日記本文・ツッコミ・トラックバックを検索できる
- http://docs.tdiary.org/ja/?rast-register.rb

pg_rast

- PostgreSQL のデータベースから全文検索を提供
- INSERT 文の実行時に Rast の検索情報データベースを更新
- http://pgfoundry.org/projects/pgrast/

mod_search_rast

- Apache 2.0 系用全文検索モジュール
- A9 OpenSearch、RSS 1.0 形式での検索結果の取得も可能
- http://module.jp/blog/mod_search_rast_0_0_2.html

デモ

http://www.unicord.org 内を検索するデモを行います

終り

ご静聴どうもありがとうございました。

- プロジェクトページ
  - http://projects.netlab.jp/rast/