世界規模ソースコード検索エンジン

―オープンソース開発の促進を目指して―

1 背景

近年、オープンソースのソフトウェア開発が活発化している。オープンソースの開発では、他のプロジェクトのソースコードを自由に閲覧し、一定のライセンスに基づいてソースコードの再利用を行うことができる。これにより、他のプロジェクトのソースコードから技術の習得を行ったり、自らのプロジェクトにコードを取り入れて再実装を避けるといった開発上の大きな利点が得られる。

しかし、これまで、他のプロジェクトのソースコードを参考にしようにも、必要なコードを探し出すのは容易ではなかった。実用的なソースコード検索エンジンは存在せず、依然として「ローカルのソースコードに対して grep をかける」といった非効率な手法が取られることが多かった。

そこで、本プロジェクトでは、こうした状況を改善し、インターネット上に存在する莫大なオープンソースの資産を最大限に活用するためのソースコード検索エンジンの開発に取り組んだ。

2 目標

本プロジェクトで開発した「ソースコード検索エンジン gonzui」はソースコードを対象とした検索エンジンである。ソースコードに特化した検索エンジンを開発するにあたっては、大規模にスケールすることは当然として、関数名などの API、コメント、ライセンス、といった要素を柔軟に扱うこと、プロジェクトの人気などを反映したランキング処理、テキストエディタや統合開発環境などとの連携を重視し、高度な実用に耐えうるものを目指した。

3 開発の内容

「ソースコード検索エンジン gonzui」は主に以下の 4 つのモジュールで構成されている。

- ソースコード検索エンジンコア
- ウェブインタフェース機能
- 運用ツール
- 言語モジュール

ここではそれぞれのモジュールの設計と実装について述べる。

３．１　ソースコード検索エンジンコア

ソースコード検索エンジンコアとは、検索エンジンの要であるインデックスの処理を行う部分である。インデックスとは、ソースコードの高速な検索に必要なデータであり、概念的には、書籍の巻末にある索引（インデックス）に近い。

検索エンジンの一般的なシステム構成図を次に示す。インデクサと呼ばれるプログラムがインデックスを作成する。ユーザはウェブブラウザなどのクライアントを用いて検索インタフェースにアクセスし、検索要求を出す。検索エンジンはインデックスを用いて高速に検索を行い、クライアントに検索結果を返す。

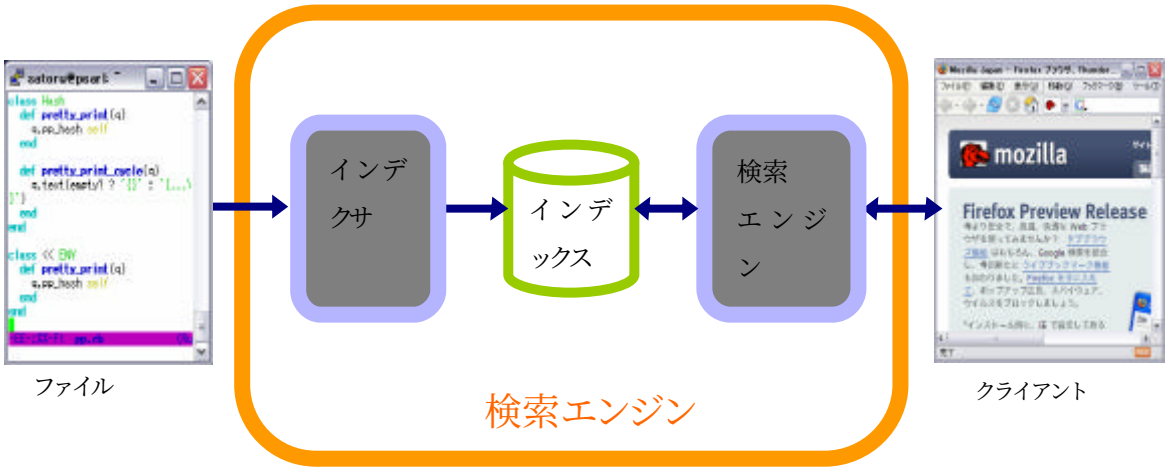

図１ 検索エンジンのシステム構成図

ソースコード検索エンジンコアでは、インデックスの作成処理、および、検索エンジンがインデックスにアクセスする処理をモジュール化している。

３．２　ウェブベース検索エンジン

「ソースコード検索エンジン gonzui」では、ウェブベースの検索エンジンを採用した。ウェブベースの検索エンジンを採用した理由を以下に挙げる。

- 多くのプラットフォームから利用できる (Windows, Mac, Linux, etc.)
- 専用のクライアントをインストールする必要がない
- Google などの検索エンジンと同様に、ブラウザから簡単に使える

gonzui ではウェブベースの検索エンジンを実装するにあたって、WEBrick という組み込み型のウェブサーバを採用した。ウェブベースのアプリケーションは一般に Apache などのサーバプロセス型のウェブサーバの下で動く CGI アプリケーションとして実装されること

が多いが、gonzui ではシステムをシンプルに構成するために、別途サーバプロセスを必要としない組み込み型ウェブサーバを採用した。

3.3 運用ツール

運用ツールとは、検索エンジンを実際に動作させるために必要なツール類のことを指す。「ソースコード検索エンジン gonzui」では以下の 3 つの運用ツールを備えている。

- gonzui-import
  ソースコードをインデックスに取り込むためのツール。zip や tar.gz などのアーカイブファイルに対応している他、Debian GNU/Linux の apt-get コマンドを用いて、ネットワーク越しのソースコードの取り込みにも対応している。
- gonzui-remove
  ソースコードをインデックスから削除するためのツール。削除されたソースコードは検索にヒットしなくなる。
- gonzui-db
  インデックスの各種情報を調査するためのツール。インデックス内のパッケージ数、ファイル数、キーワード数といった統計情報を取得できる。また、インデックスが破損していないか検証することもできる

3.4 言語モジュール

- 言語モジュールは、さまざまなプログラミング言語のソースコードを読み取るためのモジュールである。「ソースコード検索エンジン gonzui」では現在、C, C++, Java, Ruby, Python, Perl PHP などのプログラミング言語に対応している。世界的に広く使われているプログラミング言語の多くに対応しており、gonzui の適用可能範囲は広い。

4 従来の技術（または機能）との相違

gonzui が従来の検索エンジンと最も異なる点ソフトウェアのソースコードの特性を活かして検索を行える点にある。上述した言語モジュールでは、各種のプログラミング言語によって書かれたソースコードを読み取り、関数名、文字列、コメント、予約語などの情報を取り出す。ソースコード検索エンジンが従来の検索エンジンと最も異なる点は、これらソースコードに特有な情報を活かして検索を行うところにある。gonzui には次のような特徴がある。

- 大量のパッケージに対応したソースコード検索エンジンである
  大量のパッケージに対応したソースコード検索エンジンのソフトウェアは、これまでほとんど存在しなかった。本プロジェクトで開発した gonzui は、大量のパッケージの検索に対応した実用的なソースコード検索エンジンとしてユニークな存在にある。

- 主要なプログラミング言語に対応している
  gonzui は C, C++, Java をはじめとする主要なプログラミング言語に対応している。また、新しいプログラミング言語の対応はプラグイン方式で容易に追加できる仕組みとなっている。
- 簡単に使用できる
  gonzui は検索の機能や性能を追求するだけでなく、使いやすさにも多くの注意が払われたソフトウェアである。複雑な設定などを必要とせず、簡単に使いこなすことができる。
- オープンソースである
  gonzui はオープンソースのソフトウェアとして公開されている。誰でも自由に使えるだけでなく、gonzui のソースコードを読むこともできる。gonzui のライセンスは GNU Public License 2.0 を採用している。

５ 期待される効果

gonzui が普及することによって、オープンソース開発の促進が期待される。gonzui を用いることにより、オープンソース開発者は他のプロジェクトのソースコードを容易に閲覧できる。これにより、開発者の技術力の向上や、ソフトウェアコンポーネントの再利用などの効果が期待できる。

６ 普及（または活用）の見通し

現時点でも gonzui は十分に実用的な完成度に達しているが、まだまだ発展性のあるソフトウェアである。今後も開発を継続し、新しいプログラミング言語の対応や高速化、国際的認知度の強化などに取り組んでいきたい。

７ 開発者名

メイン開発者
- 高林哲 (satoru@namazu.org)

協力開発者
- 田中哲 (akr@m17n.org)
- 西田圭介 (knishida@open-cobol.org)