D１

ソフトウェアプロダクトの収集・解析・検索システム

大阪大学　大学院情報科学研究科　○井上　克郎

Software Product Archiving and Retrieving System

Katsuro Inoue, Graduate School of Information Science and Technology, Osaka University

Abstract:
Collections of already developed programs are important resources for efficient development of reliable software systems. In this research, we propose a novel model of ranking software components, called component rank, based on analyzing actual use relations among the components and propagating the significance through the use relations. We have developed a Java class search system named SPARS-J using the component rank, and applied SPARS-J to various collections of Java programs. The result shows that SPARS-J ranks generic and more specific components higher. SPARS-J has also been applied to several companies, and it showed its capability of searching components in the company assets.

１. はじめに

これまでに、様々なソフトウェアが数多く開発され、世界中で利用されてきている。しかし、開発されたソフトウェアが、有効に蓄積され、再利用されているとは言い難い状況である。本研究では、ソフトウェアを対象とした検索システムの構築を目指す。まず、世界中で開発され、公開されているソフトウェアのソースコード、オブジェクトコード、ライブラリ、ミドルウェアやそれらに関するドキュメント等のソフトウェアプロダクトの自動収集を行う。その際、高速ネットワークが必要である。次に、収集したプロダクトを解析モデルに従って解析、分類し、ソフトウェアプロダクトアーカイブを構築する。ユーザは高速ネットワークを通じてこのアーカイブに対し、システムの仕様や機能、ライブラリ名、プログラムパターンなど種々の問合せを与えることにより、類似のシステムやライブラリ、プログラムのパターンなど関連するソフトウェアプロダクトの情報を容易に入手することが期待できる。

２. 研究開発項目とその成果概要

２. １ 概要

本研究では、ソフトウェアを対象として、Webページの検索エンジンとしてよく知られているgoogleのような検索システムの構築を目指した。具体的には、まず、世界中で開発され、公開されているソフトウェアやそれに関するドキュメント等のソフトウェアプロダクトの自動収集を行う。次に、収集したプロダクトを解析モデルに従って解析・分類し、ソフトウェアプロダクトアーカイブを構築する。ユーザは高速ネットワークを通じてこのアーカイブに対し、システムの仕様や機能、ライブラリ名、プログラムパターンなど種々の問合せを与えることにより、類似のシステムやライブラリ、プログラムのパターンなど関連するソフトウェアプロダクトの情報を容易に入手することができる。

研究遂行にあたっては、以下の６班でそれぞれサブテーマを設定した。

(1) 第１班（総括班）：ソフトウェアプロダクトの収集・解析・検索システム

対象ソフトウェアプロダクトをJavaに限定し、Java部品検索システムSPARS-Jを実装した。インターネット上からJavaの部品(クラスファイル)を収集し、部品間の利用関係を解析することにより部品の順位付けを行い、ユーザから入力されたキーワードに一致する部品を順位に基づいて出力するシステムである。

(2) 第２班：イディオムを用いたソフトウェア検索方式の研究

ソフトウェアの意味的検索をソフトウェア中のイディオムの検索で実現する方式について研究を行った。

(3) 第３班　バイナリ形式コンポーネントの収集・解析・検索システムの開発

バイナリ形式部品からパターンを検索するアルゴリズムを開発し、プロトタイプシステムを実装した。また、検索エンジンに与えるクエリのユーザビリティや、パターン類似度に基づくランク付けの妥当性について検討した。

(4) 第４班：ソフトウェア再利用のための分散作業支援方式の研究

セマンティックWebをソフトウェア再利用の情報提供の枠組みとして利用する方法について提案した。セマンティックWebとはWeb文書にメタデータを付与し、文書間の意味的なつながりを表そうとする技術である。この意味情報を利用することによって、求めているソフトウェアを的確に検索できるシステムの構築方法を検討した。

(5) 第５班：ソフトウェアメトリクスに基づく中粒度解析システムの開発

主に、インターネット上からソフトウェアプロダクトデータを自動的に収集する「巡回ロボット」の開発と収集データから必要なプロダクトを検索するためのクエリについて検討した。

(6) 第６班：認知モデルに基づいた関連ソフトウェア自動収集提示システムの開発

仕事に有益なソフトウェアプロダクトを収集するために、プログラミング環境における作業から認知的タスクモデルを構築し、自動的に関連するソフトウェアプロダクトを収集しユーザに提示する支援の枠組みを構築し、それに基づくシステムを開発した。

これらの成果は、学術論文12本、特許2本としてまとめられている。また、本テーマに関するワークショップを２回開催している。

2.2 Javaコンポーネント収集・分析・検索システムSPARS-J

ここでは、第一班の成果についてまとめる。近年多くのソフトウェア開発で用いられるようになったプログラミング言語 Javaを対象としたJavaコンポーネント収集・分析・検索システムSPARS-Jの構築を行った。 SPARS-Jの概要を図１に示す。SPARS-J はJava ソフトウェア部品の容易な検索、参照を実現するためのシステムである。キーワードを検索キーとして、検索キーと関連したソフトウェアのソースコードを効率良く検索することが可能である。さらに、検索結果表示の際にソフトウェア部品に関する詳細な情報を併せて提供する。なお、SPARS-Jにおけるソフトウェア部品とは、Javaのクラスまた

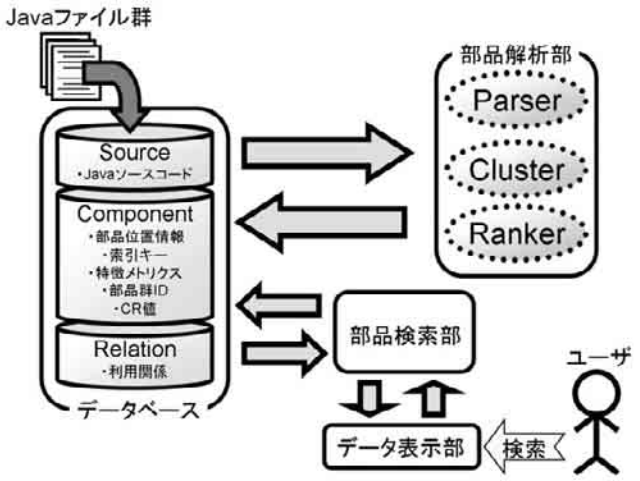

図１　SPARS-Jシステム概要

はインタフェースを単位とするソースコードを指す。

以降、SPARS-Jシステムの主な機能、構成、特徴、評価についてまとめる。

SPARS-Jシステムの主な機能

(1) キーワード検索: ユーザが要求したキーワードを検索キーとして部品の検索を行なう。
(2) 部品間の利用関係の提示:実際の部品の使用方法に関する情報を得ることが可能になる。
(3) パッケージブラウザ: クラス階層構造を表示し、パッケージ内に含まれるクラスの一覧を取得できる。

SPARS-Jシステムの構成

(1) データベース
- Sourceリポジトリ：Javaで記述されたソースコードのファイルを保存する。
- Componentリポジトリ：Sourceリポジトリ中のJavaファイルの解析で得られた部品に関する情報を格納する。
- Relationリポジトリ：各部品間の利用関係を格納する。

(2) 部品解析部

Javaファイルを解析し、検索時に必要な情報を抽出し、各種リポジトリに格納する。
- Parser：Javaファイルの構文解析を行い、部品の取り出し、特徴メトリクスの計測等、検索に必要な情報を抽出する。
- Cluster：Parserで求めた特徴メトリクスを用いて類似部品を部品群にまとめる。
- Ranker：Relationリポジトリの情報から、 Component Rank(CR)法(部品の利用関係からソフトウェア部品の利用実績を測定し、利用実績が多い部品ほど重要度が高くなる部品評価手法)に基づいて、各部品のCR値の計算を行う。

(3) 部品検索部

ユーザによって与えられた検索条件から検索の種類を判別し、要求と一致する部品を検索する。また、検索結果の部品を評価値(CR値とTF-IDF値(任意の部品中における特定の単語の出現頻度、および、特定の単語を含む部品数の逆数の値を正規化して部品を重み付けした値）の合成値)が高い順に順位付けして、データ表示部に渡す。

(4) データ表示部

部品検索部とユーザを結ぶWebインタフェースである。検索の結果ヒットした部品を部品リストの順位で表示する。

SPARS-Jシステムの特徴

(1) キーワード型検索

複数のキーワード（例えば、プログラム中のコメント、変数、メソド名）を入力するだけで、高速に対応する部品が、部品の有用度に従って順位付けされ表示される。

(2) 部品順位付け手法

部品の順位付けには、CR法とTF-IDF法を併用した手法を用いている。これにより、よりユーザのクエリに合致し、重要度の高い部品が検索できる。

(3) 類似ソフトウェア部品の統合

部品の中にはコピーされた部品やコピーされ一部だけ変更された部品のように、類似した部品が数多く存在している。SPARS-Jでは類似した部品をまとめて取り扱うことにより、より正確に利用実績が測定できる。

(4) 再利用時に有益な情報の表示

単に部品名や部品のソースコードを表示するだけでなく、部品間の利用関係（検索された部品が利用している部品、あるいは、利用されている部品)、同一部品群に属する部品、パッケージブラウザによるクラス階層構造、等の再利用時に有益な情報を提示できる。

SPARS-Jシステムの評価

インターネット上から収集した約16万個のJavaクラスをSPARS-Jシステムを用いて解析し、検索実験を行った。その結果、ユーザが期待するJava部品の情報が効率よく検索できることが確認できた。

3. ネットワークの活用について

主に、インターネット上からのソフトウェアプロダクト収集にネットワークを利用した。特に、第５班は、インターネット上からソフトウェアプロダクトを自動収集する巡回ロボットの開発を行っており、ネットワークを最大限に活用している。ここでは、その結果について簡単にまとめる。

ネットワーク上での巡回ロボットの検索効率を上げるため、ソフトウェア資源を扱っていると予想されるウェブコミュニティを既存のWWW検索エンジンを用いて発見し、そのコミュニティのURLを巡回の起点とする。但し、巡回対象としたウェブコミュニティによって、含まれるソフトウェア資源の量には大きなばらつきがあることが分かった。そこで、起点URLからの巡回数（巡回URL数）を x とし、$n_{i-1} \leqq x \leqq n_i$（iは自然数）においてソフトウェア資源が m 個以上発見された場合にのみ、$n_i \leqq x \leqq n_{i+1}$の巡回を行うこととした。この結果、プロトタイプシステムの巡回ロボットを24時間連続稼動させた結果、クラス約2,000個、メソッド約35,000個を検索することができた。なお、検索されたプログラム資源には、既存のWWW検索エンジンのみでは検索が非常に困難と思われる資源も多数含まれていたが、市販ソフトウェアのアップデートファイル、ソースコードが含まれない圧縮ファイルなど、有効でない資源も15%〜30%含まれていた。有効でない資源の割合は、設定する起点URLに大きく左右されることが確かめられており、ソフトウェア資源が発見されたページの特徴的な単語をキーワードとして起点URLを検索するなど、更なる改良の余地がある。

4. まとめ

対象ソフトウェアプロダクトをJavaに限定し、Java部品検索システムSPARS-Jを実装した。SPARS-Jは、インターネット上からJavaの部品（クラスファイル）を収集し、部品間の利用関係を解析することにより部品の順位付けを行い、ユーザから入力されたキーワードに一致する部品を順位に基づいて出力するシステムである。

インターネット上から収集した約16万個のJavaクラスをSPARS-Jシステムを用いて解析し、検索実験を行った。その結果、ユーザが期待するJava部品の情報が効率良く検索できることが確認できた。既に、我々の研究成果を知った幾つかのソフトウェア組織から、ソフトウェア部品の再利用性の評価、あるいは、再利用部品検索システムとして実際の開発現場で利用したいという問い合わせも来ている。実際の開発現場への適用を通じて、SPARS-Jシステムの改良、応用を今後とも続けていく予定である。

5. 研究開発実施体制

代表研究者　大阪大学　大学院情報科学研究科　井上克郎

研究分担

研究開発項目：ソフトウェアプロダクトの収集・解析・検索システム

大阪大学　大学院情報科学研究科　井上克郎

大阪大学　大学院情報科学研究科　楠本真二

大阪大学　大学院情報科学研究科　松下　誠

立命館大学　情報理工学部　山本哲男

研究開発項目：イディオムを用いたソフトウェア検索方式の研究

名古屋大学　大学院工学研究科　阿草清滋

愛知県立大学　情報科学部　山本晋一郎

研究開発項目：バイナリ形式コンポーネントの収集・解析・検索システムの開発

奈良先端科学技術大学院大学　情報科学センター　飯田元

研究開発項目：ソフトウェア再利用のための分散作業支援方式の研究

神戸大学　大学院自然科学研究科　荻原剛志

研究開発項目：ソフトウェアメトリクスに基づく中粒度解析システムの開発

奈良先端科学技術大学院大学　情報科学研究科　松本健一

奈良先端科学技術大学院大学　情報科学研究科　門田暁人

研究開発項目：認知モデルに基づいた関連ソフトウェア自動収集提示システムの開発

東京大学　先端科学技術研究センター　中小路久美代

東京大学　先端科学技術研究センター　山田和明

University of Colorado　Yunwen Ye