6B-3

API の類似性を利用したソフトウェア類似部品検索手法の評価

高見 愛† 井上 勝行† 北村 操代†

三菱電機 先端技術総合研究所†

1. はじめに

企業のソフトウェア開発では、過去の開発における成果物を再利用して効率化を図ることが求められる。成果物の再利用可能な単位をソフトウェア部品として管理しておけば、新たな開発で流用もしくは開発の参考にすることができる。しかし、管理されている大量の部品から、所望の部品を検索することは容易ではない。

筆者らは、ソフトウェア部品の公開 API の類似性を利用して、部品の提供する機能が似ている部品を検出する方法を提案している[1]。これにより、キーワードで関連付けられた部品だけでなく機能の類似する部品も検索結果に含めることができるようになる。

本稿では、提案手法と自社製 C++クラスライブラリを対象とした評価について述べる。

2. ソフトウェア部品間の類似度

2.1 既存の類似度算出手法

ソフトウェアの類似度を算出する既存手法には、プログラムの構造の類似性[2]や、置換え時の修正コスト[3]を利用する方法がある。プログラムの構造を利用する方法は、部品の実装方法に強く依存しており、コピー部品の検出に適しているが、異なる開発者が作成した部品を抽出できない。また、置換に必要な修正コストを利用する方法は、少ない作業量で置換可能な部品を検出できるが、機能の一部が似ている部品の候補を抽出するには厳密すぎる。

2.2 提案手法

提案手法では、部品の公開 API に着目し、API 中の型や識別子が似ていれば API も似た機能をもつと仮定し、API の型と識別子の類似性を用いてソフトウェア部品間の類似度を算出する。具体的には、型、識別子が同一または類似する場合に類似度の得点を加算し、最終的な得点が閾値を超えると API が類似すると判定する。さらに、ソフトウェア部品に対して、その部品の公開 API に類似する API を多くもつ部品を類似部品と判定する。これにより、同一の API でなくても型や識別子が似ていれば類似と見なすことができるので、機能の一部が似ている部品も広く抽出できるようになる。

Evaluation of Retrieving Method of Similar Software Components using Similarity of API.

† Advanced Technology R&D Center, Mitsubishi Electric Corp.

API の類似性は、API の型および、API の識別子の類似性から求める。型は、関数の戻り値の型、変数の型、引数の型を含む。型が同一もしくは互換である場合、API の入出力となるデータが似ていると見なして、API は似た機能を有すると判定する。識別子は、関数の名前と変数の名前、引数の名前を含む。識別子が似ていれば、API は似た機能を有すると見なす。識別子の類似判定には、類語辞書を入力として与え、類語の場合に類似すると見なし、文字単位での比較は行なわない。

API の類似度は、型に関する類似度の得点と、識別子に関する類似度の得点を足し合わせ、取り得る得点の最大値で割ったものとする。API の類似度が予め定めた値(API 間類似度判定閾値)を超えたとき、API は類似と見なす。

ソフトウェア部品の類似性は、類似する API の割合によって判定する。具体的には、部品 P の部品 Q に対する類似度は、部品の公開 API のうち、部品 Q に類似する API が存在するものの割合とする。この割合が、予め定めた値(部品間類似度判定閾値)を超えたとき、部品 P は部品 Q に類似すると見なす。

2.3 ソフトウェア類似度算出ツール Collate++

提案した類似度の算出方法に基づき、C++言語で開発されたソフトウェア部品の類似度を算出するツール Collate++を実装した。

Collate++はヘッダファイルに宣言される API のうち、public および protected 宣言されたメンバ変数とメンバ関数を利用する。

型の類似度は、型の互換性に基づき、0～5 点に採点する。int や char など、組込み型の類似度は予めユーザが定義したものを利用する。型がソースコード中に宣言されたクラスである場合は、クラスの継承関係を元に類似度を求める。

識別子の類似度は、二つの識別子が類語関係にあるかどうかで判定する。識別子の類似度の得点は、識別子が同一である場合に 2 点、類似である場合に 1 点、それ以外は 0 点とする。二つの識別子が類語関係にあるかは、類語辞書ファイルで類語として記載されているかどうかで判断する。Collate++では、21791 組の類語を記録した類語辞書ファイルを利用する。

3. 既存ライブラリを対象とした評価

3.1 評価実験

提案手法の評価として、提案手法で算出した類似度の妥当性、API 類似度算出における各パラメータの影響度を調査する。

類似度の妥当性では、評価対象ライブラリの開発に携わる熟練者 1 名の判断結果との差異を調査する。熟練者は提案手法により抽出した類似部品が、流用できる、もしくは参考にできる場合に似ていると判断する。また、類似度算出方法の詳細を評価するため、API 間類似度のパラメータである型と識別子、引数の類似度のうち、いずれの要素が支配的になるかを分析する。

本実験で、類似部品は、二つのライブラリから双方向に抽出する。評価には、自社製の 2 種類の C++クラスライブラリを用いる。それぞれ、ライブラリ A、ライブラリ B と呼ぶ。両ライブラリの規模を表 1に示す。双方向の抽出では、ライブラリ A のクラス Ai(計 10 個。以下、選出クラスと呼ぶ)を対象として、ライブラリ B から、クラス Ai に類似するクラス(類似クラス)と、クラス Ai に類似するクラス(被類似クラス)をそれぞれ上位 10 個抽出する。そして、これら最大 20 個の部品を類似部品とする。

なお、ソフトウェア部品の単位は、クラスまたは構造体とする。事前調査により、API 間類似度判定閾値と部品間類似度判定閾値は共に 0.5 とする。また、API 数が 10 未満のクラスは評価から除外する。

表 1 対象ライブラリの規模

	ライブラリ A	ライブラリ B
ソースファイル	115 個	200 個
うちヘッダファイル	59 個	102 個
実行数	33k 行	83k 行

3.2 実験結果

全ての選出クラスについて熟練者による判断結果を図 1に示す。熟練者が似ている判断した類似部品の 8 割以上が直接の継承関係がないものだった。また、算出パラメータの影響度を図 2のグラフに示す。図は、API 間類似度が閾値以上だったときに、型と識別子、引数の類似度が合算後の値の過半数を超えて支配的になる割合を示している。全体の 6 割以上で、識別子の類似度が API 間類似度の過半数を超えることがわかる。

4. 考察

熟練者による判断結果より、全ての選出クラスについて提案手法が抽出した類似部品の中に熟練者が似ていると判断する部品が含まれることがわかる。類似部品には、熟練者が似ていないと判断した部品も含まれるため、全てが有効ではないが熟練者の知識なしでも類似する部品の候補を広く抽出するには有効といえる。また、類似部品の多くが継承関係の強くない部品であったことから、継承関係を辿るだけでは見つけにくい類似部品も抽出できることがわかる。

また、算出パラメータの影響度を分析した結果、識別子の類似度が大きな値をとることがわかった。Collate++では、内部処理でメンバ変数および関数の識別子にクラス名を付与するため、クラス名が似ているだけで識別子が高くなりやすいことが要因の一因と考えられる。

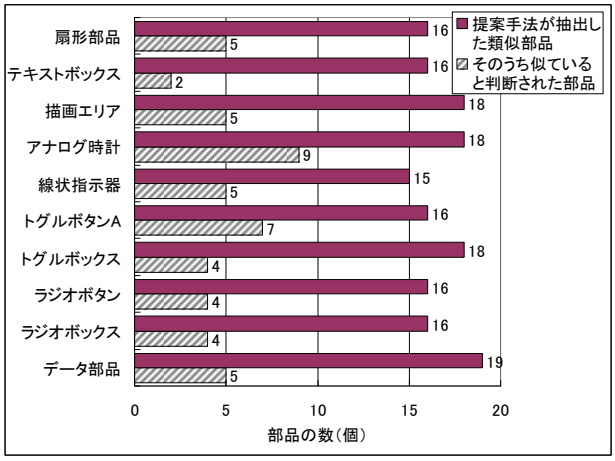

図 1 熟練者による判断結果

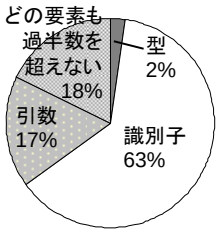

図 2 API 間類似度への影響度

5. まとめ

本稿では、製品に適用されているライブラリを用いて、提案手法の評価を行なった。熟練者が似ていると判断したクラスの大部分を、提案手法が類似部品として抽出できることを示唆する結果を得た。今後は、類似度の算出パラメータの重み付けを検討するなど、評価を行ないながら、算出方法の改善を図っていく。

参考文献

1) 高見, 他：ソフトウェア部品検索のための API を利用した類似部品検出方法の提案, 情報処理学会研究報告, Vol.2012-SE-176, No.6(2012).
2) 山本, 他：ソフトウェアシステムの類似度とその計測ツール SMMT, 電子情報通信学会, Vol.J85-D-I, No.6, pp.503-511(2002).
3) 鷲崎, 他：有向置換性類似度に基づくコンポーネント検索方式の実現と評価, 情報処理, Vol.43, No.6, pp.1638-1652(2002).