大阪大学
平成 14 年 5 月 2 日

高速コードクローン検出システムの開発

- プログラムの中に存在する同形部分（コードクローン）を高速に検出
- 柔軟に多様なプログラミング言語に対応
- 大規模なプログラムの類似度や進化度を客観的に測定

大阪大学大学院情報科学研究科コンピュータサイエンス専攻ソフトウェア工学講座 （井上克郎教授）では、大規模なプログラムの中に存在するコードクローンを高速に検出するシステムを開発した。このシステムは、対象とするプログラムを限定された語の列に変換した後に、同形列検出アルゴリズムを適用することにより、多少の差異があるコードクローンも高速に検出できる。また、C や Java、COBOL など多様なプログラミング言語に容易に対応させることができる。本システムの完成により、大規模なプログラムの間のコードクローンが検出できるようになり、大規模なプログラムの類似度を客観的に計測できる。今後、プログラムの保守や著作権の管理に応用することが期待される。本システムの内容は、電気電子学会ソフトウェア工学論文誌(IEEE Transactions on Software Engineering)6 月号に掲載される。

□ 背景

何度も改良を加えながら長期間使われ続けるプログラムが世の中には多数存在する。そのようなプログラムの改良の際、プログラムの断片を複製し、一部改変して別の部分で利用することがよく行われる。このようにして、プログラムにコードクローンが含まれるようになるが、プログラムが大規模になると、人間の手でコードクローンを追跡することはほとんど不可能である。今までにもコードクローンを検出するシステムはあったが、50 万行程度のプログラムの解析に 6 時間ほど要するなど効率が悪い、改行位置やコメントの変更、変数名の付け替えなど多少の変更を加えられたコードクローンを発見できないなど、実用性で問題があった。

□ 開発したシステム

対象とするプログラムから、コメントや改行記号などを削除し、語の切れ目で分解し、語の列に変換する。その際、対象とするプログラミング言語特有の変換ルールを用いて、例えばテーブルの初期化など、コードクローンとして意味のないものを削除する。

その後、高速な同形列検出アルゴリズムを用いて、同一の列を検索し、コードクローンとして出力する。この手法により、例えば 1000 万行のプログラムの解析をパソコンで 1 時間程度でできるようになった。また、C, C++, Java, COBOL, LISP など多様な言語のプログラムに対応できるようになった。

□　用途

○プログラムの保守

プログラムを変更する場合に、コードクローンがあると、そのすべてを変更する必要がある。従って、まず、このシステムでコードクローンを発見しておき、変更が必要な箇所を探しておく。さらに、コードクローン部分を一箇所にまとめて、プログラムを再構成すると、保守のしやすいプログラムに改良することができる。

実施例 1：20 年以上、改良を重ねて使われてきている公共システムのプログラムの解析を行った。プログラムの開発者が気がつかずにいた多数のコードクローンを発見し、変更作業の指針を与えた。

○プログラム間の類似コードの発見

同じ祖先を持つプログラムで、別の改良を加えられて現在に至る、複数のプログラムが、どの程度似通っているか、違っているかなどを客観的に調べることができるようになる。

実施例 2：3 種類の UNIX(Linux, FreeBSD, NetBSD)の間のコードクローンの検出を行ない、FreeBSD と NetBSD が類似しており、Linux がそれらから独立していることを定量的に確かめた。

実施例 3：学生が演習で作成したプログラムの間のコードクローンを検出し、不正コピーを発見した。

実施例 4：プログラムの不正盗用の民事裁判において、2 つのプログラムの差異の度合を定量化し、証拠資料として提出した。

□　用語

コードクローン

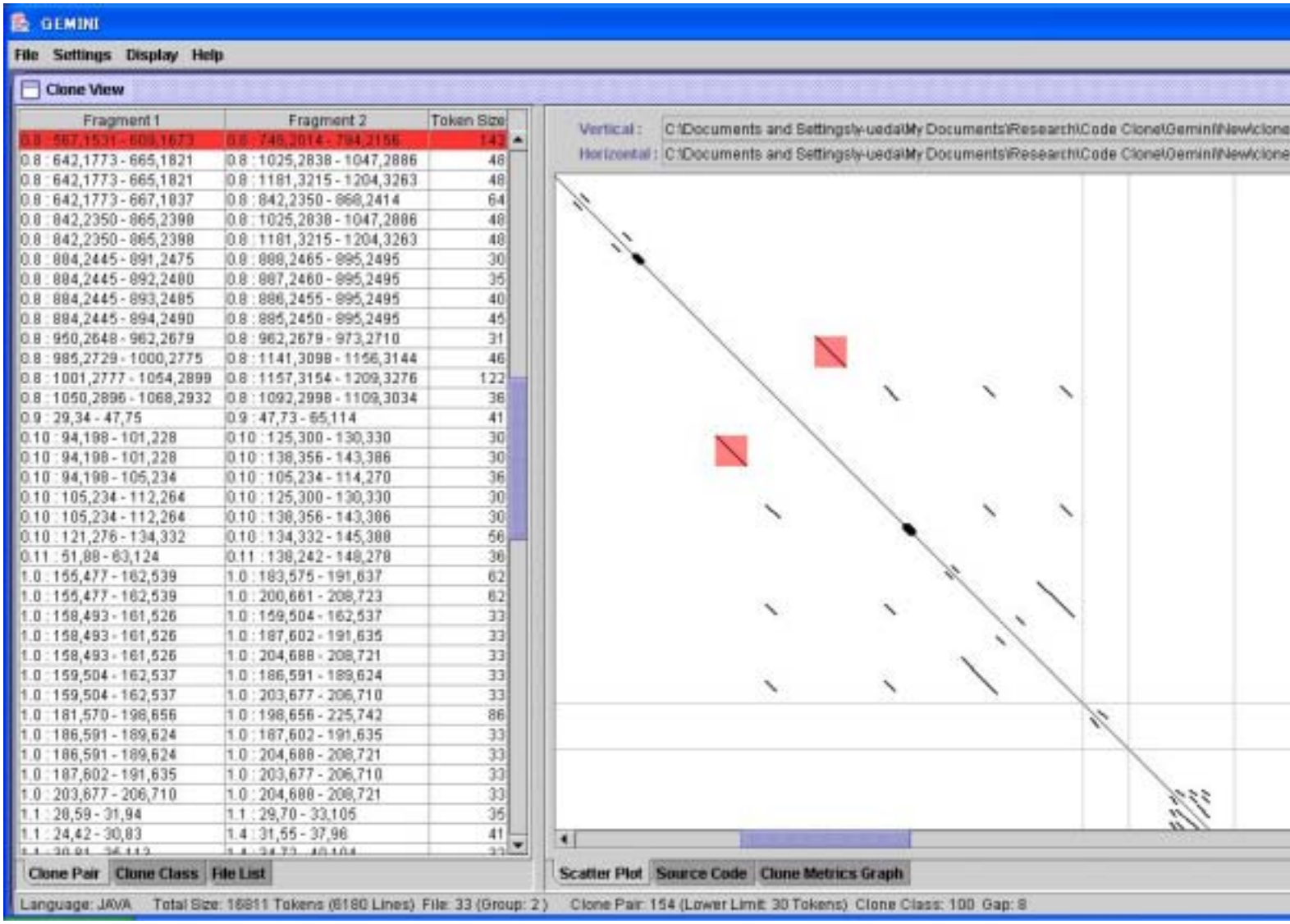
一対のプログラムの断片で、他の場所に、ほぼ同じ形の断片が存在するもの。プログラムの開発や改良の際に、一部のプログラムの断片をコピー・ペーストしてそのまま利用したり、変数名やコメントの変更など軽微な変更を施して利用した場合などに生じる。それらの断片にバグが存在した場合、全てのコードクローンの修正を要するなど、コードクローンがあるとプログラムの保守が困難になる。

電気電子学会ソフトウェア工学論文誌(IEEE Transactions on Software　Engineering)

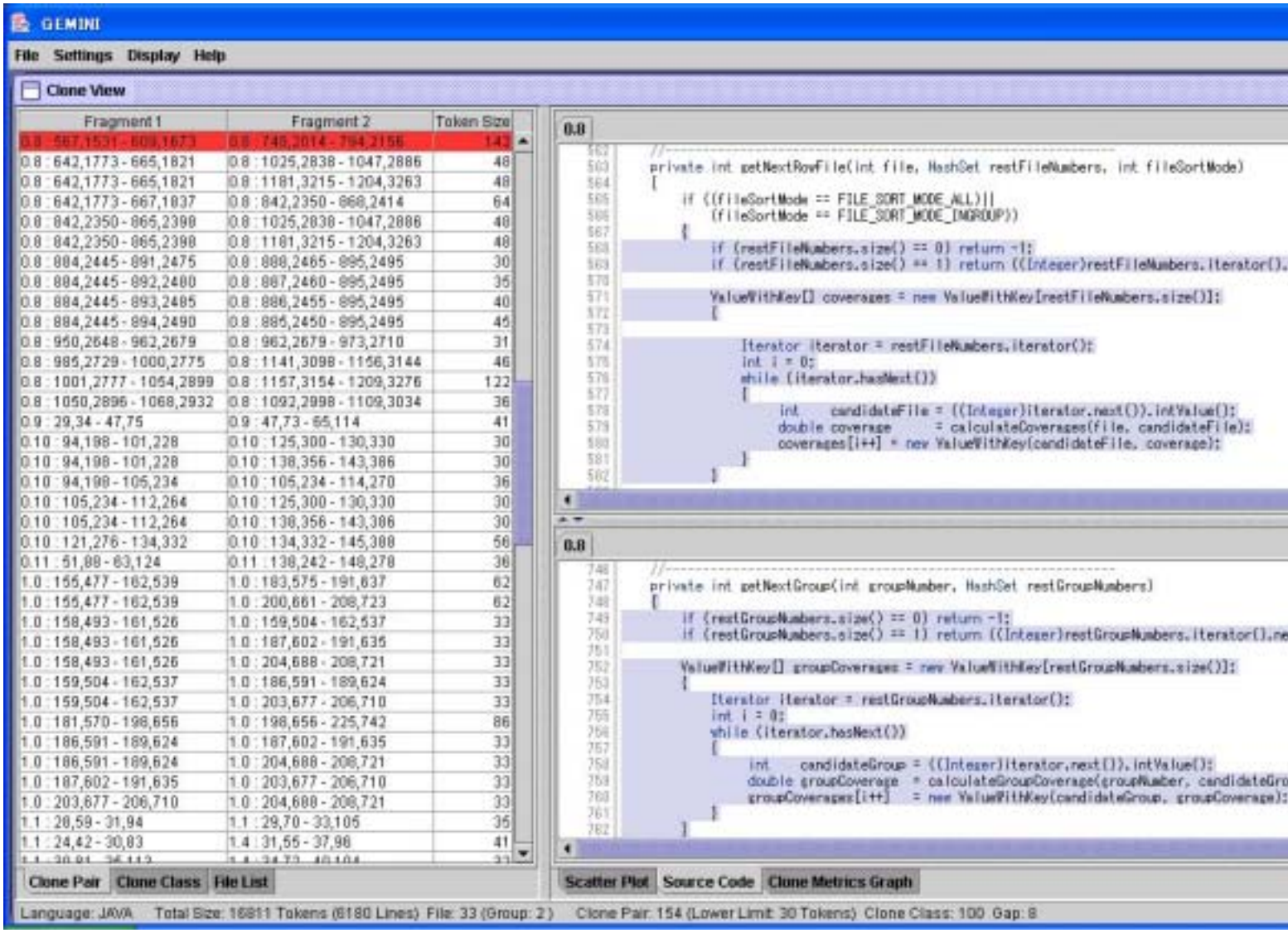
電気電子学会（IEEE:　The Institute of Electrical and Electronics　Engineers, Inc.）

は、世界各国の会員約 37 万 7 千人からなる情報、電気、電子、バイオ、医療などの分野を対象とした学会で、そのソフトウェア工学論文誌にはコンピュータのソフトウェアに関する第一級の論文が毎月 5〜10 件程度掲載される。

□　本システムの出力例



上図の右の窓は、本システムの出力例。２つのプログラムの語がそれぞれ X 軸と Y 軸に展開され、同一のもののところに黒点が打たれる。右下に伸びる線分がそれぞれコードクローンを示している。対角線上は常に同一なので直線が引かれ、また、対角線を軸として線対称になる。赤い部分が今、注目しているコードクローンを示す。左の窓はプログラム中のコードクローンの位置を示している。



上図の右の窓は、発見したコードクローンを表示した例で、上下の網掛けの部分が対応する一対のコードクローンになる。

問い合わせ先

大阪大学大学院情報科学研究科コンピュータサイエンス専攻
ソフトウェア工学講座　教授　井上克郎
〒560-8531　大阪府豊中市待兼山町１－３
TEL: 06-6850-6570
FAX: 06-6850-6574
EMAIL inoue@ist.osaka-u.ac.jp