

Prediction and Ranking of Co-change Candidates for Clones

Author :
Manishankar Mondal
Chanchal K. Roy
Kevin A. Schneider

MSR 2014 Proceedings of the
11th Working Conference on
Mining Software Repositories
Pages 32-41

目次

- 背景, 目的
- アプローチ
- 分析結果
- まとめ, 所感

背景，目的

背景

- コードクローンを変更する必要があるが生じたとき, どのようなクローンを変更すべきかを考える必要がある
- また, どのようなクローンを優先して変更すべきかも考える必要がある
- 以上の2つのことを考えながらクローンに対処する必要があり, 効率的でない

目的

- どのようなコードクローンを共変化すべきかを予測する
- どのようなコードクローンを優先して変更すべきかを考える

アプローチ

分析の方針

- OSSのVCSの情報を用いて予測手法を構築する
- evolutionary coupling (EVと略す) [8]を用いて予測を行う
 - EV: ソフトウェアの進化の中で2つ以上のプログラムの要素(ファイル, クラス, メソッドなど)がしばしば同じように変化すること
- 共変化するクローン片のペアをPair of Co-canged Clone Fragments (PCCFと略す)とし, EVを示すクローン片として扱う

[8] H. Gall, K. Hajek, and M. Jazayeri, "Detection of logical coupling based on product release history," Proc. ICSM, 1998, pp. 190–199.

リサーチクエスト

- RQ1: EVを用いて特定のクローン片に対して共変化の候補を予測することが出来るか
- RQ2: コードの変更の頻度よりどれだけ最近変更されたかを考慮することで, EVを示す共変化の候補のランク付けの精度は良くなるか
- RQ3: EVを示すクローン片の特徴はどのようなもので, その中でどのような特徴がEVを示さない共変化の候補のランク付けをより良くすることが出来るか
- RQ4: EVを示す候補と特定のクローン片でEVを示さない候補の両方のタイプをランク付けするにはどうすればよいか

分析結果

データセット

Table 2: Subject Systems

Sys.	Lang.	Domains	LOC	Revs
Ctags	C	Code Def. Generator	33,270	774
QMail Admin	C	Mail Management	4,054	317
jEdit	Java	Text Editor	191,804	4000
Freecol	Java	Game	91,626	1950
Carol	Java	Game	25,091	1700
Jabref	Java	Reference Manager	45,515	1545
Revs = Revisions.		Sys = Systems		

分析準備

1. 全リビジョンを取り出す
2. それぞれのリビジョンから全メソッドを取り出す
3. リビジョン間で各メソッドを系譜でつなぐ
4. NiCad[4]を用いてメソッドの系譜の中からコードクローンを検出する
5. クローンであるメソッドの系譜のリビジョン間の変更を取り出す

[4] J .R. Cordy and C.K. Roy, “The NiCad Clone Detector”, Proc ICPC Tool Demo, 2011, pp. 219 – 220.

EVを示したクローン片

変更があったク
ローンの内31.85%
のクローンがEVを
持って変更された
NCEC/NCRC

全体のクローンの
内70.82%が1度も
変更されていない
1-NCRC/NC

Table 3: Statistics Regarding Evolutionary Coupling among Clones

Systems	NC	NCRC	NCEC	NPCCF
Ctags	694	412	89	79
QMailAdmin	137	128	34	347
jEdit	12329	1356	326	324
Freecol	2265	1659	489	577
Carol	3040	1365	616	2041
Jabref	3708	1552	509	684

NC = No. of Clones created during system evolution.

NCRC = No. of Clones that Received Changes.

NCEC = No. of Clones that showed Evolutionary Coupling.

NPCCF = No. of the pairs of co-changed clone fragments.

共変化の候補

- 共変化の候補となるクローン片を以下の4つに分類する

- Possible Co-change Candidates

- クローンセットの中で初めに変更されたクローン片以外のクローン片

- True Co-change Candidates

- 実際に共変化したクローン片

- Predicted Co-change Candidates

- EVを用いて予測された共変化の候補となるクローン片

- Correctly Predicted Co-change Candidates

- True Co-change CandidatesとPredicted Co-change Candidatesの積集合。予測が当たったクローン片

RQ1の予測手法

- コミットCでCFと共変化するクローン片を予測する
- CF以外のクローン片をPossible Co-change Candidatesとする
- これまでのクローン片の変更履歴から得たEVからPredicted Co-change Candidatesを選ぶ
- コミットCからTrue Co-change Candidatesを得る
- Predicted Co-change CandidatesとTrue Co-change CandidatesからCorrectly Predicted Co-change Candidatesを得る

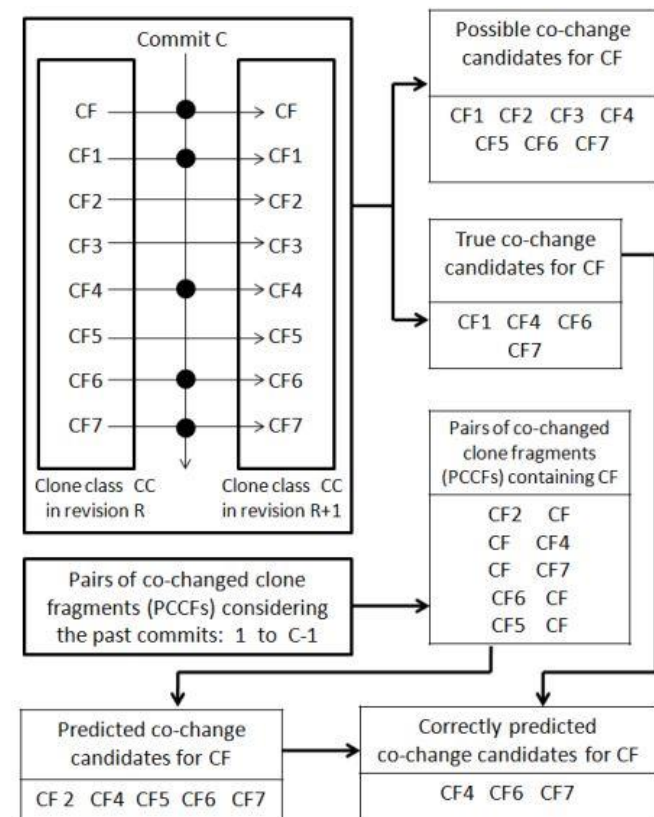


Figure 3: Prediction of co-change candidates using evolutionary coupling

RQ1の予測手法の精度

- 予測手法の精度を測るために以下の割合を計算する
 1. Possible Co-change Candidatesの中でEVを示すもの
 2. Possible Co-change Candidatesの中でEVを示さないもの
 3. EVを示すものから予測したものの内実際に共変化したものを当てた割合 (適合率)
 4. 予測していない候補から実際に共変化したクローン片の割合
 5. 実際に共変化したクローン片の中で予測できたものの割合. この予測手法の再現率とも言える (再現率)

RQ1の予測手法の精度

- 右図は1と2の割合を示したもの

- EVを示さないクローン片が多いことがわかる

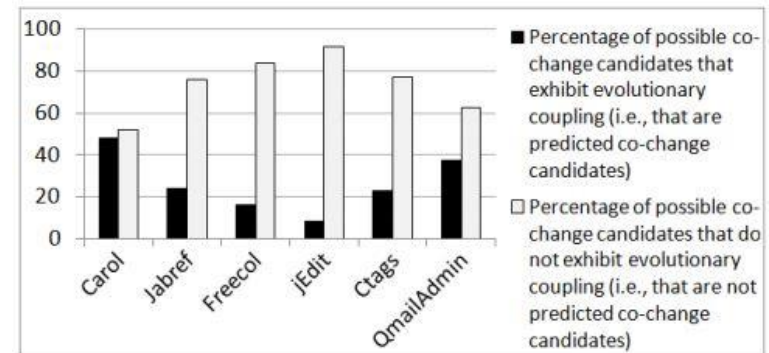


Figure 4: Comparison between the proportions of predicted co-change candidates and non-predicted co-change candidates

RQ1の予測手法の精度

- 右図は3と4の割合を示したもの

- 適合率が高いことがわかる

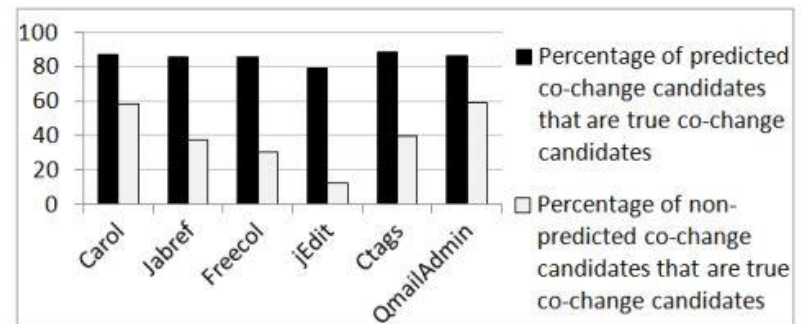


Figure 5: Comparison between the percentage of predicted co-change candidates that were true co-change candidates (i.e., the precision) and the percentage of non-predicted co-change candidates that were true co-change candidates

RQ1の予測手法の精度

- 右図は5の割合を示したもの
- 再現率はあまり高いことがわかる

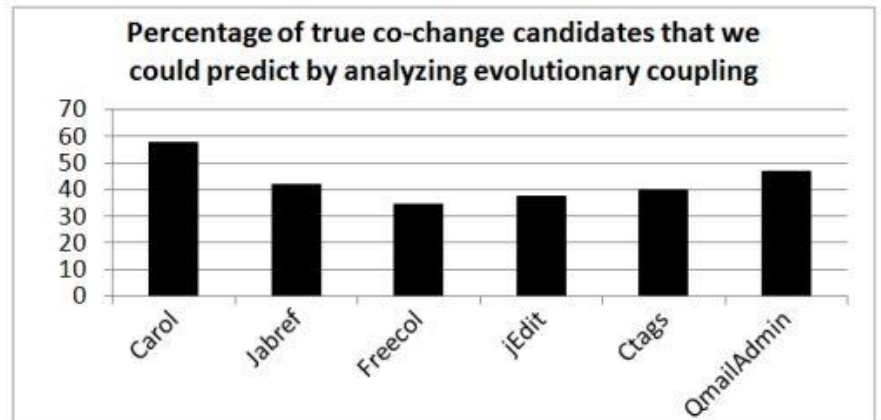


Figure 6: The proportion of true co-change candidates that we could predict by analyzing evolutionary coupling (i.e., the recall)

RQ2のランク付けの手法

- 変更の頻度かどれだけ最近変更されたかのどちらをランク付けの基準とすればランク付けの精度がよいかを調べる

- 得られたランクの和が小さい方が精度がよいとする

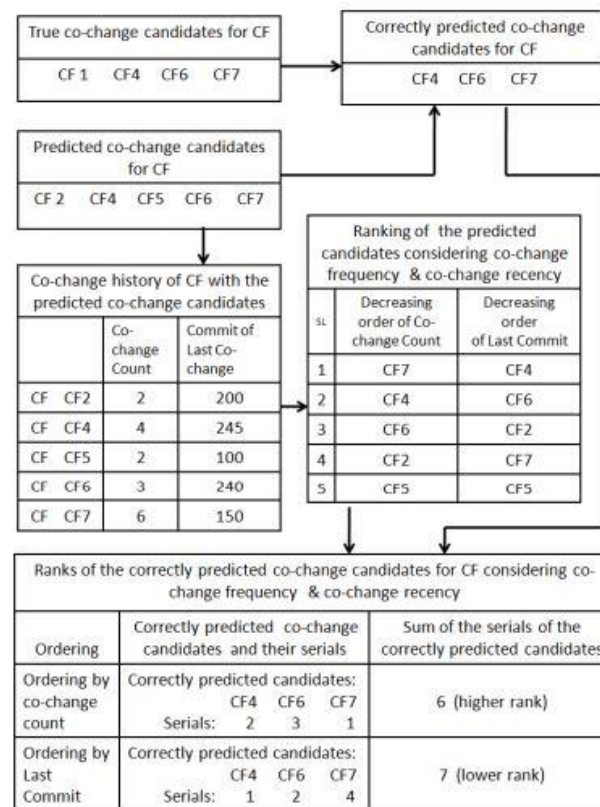


Figure 7: Ranking of predicted co-change candidates by co-change frequency and co-change recency

ランク付けの精度

- 右図は変更の頻度とどれだけ最近変更されたかのどちらをランク付けの基準にすることで精度が上がるかを比較している
- どれだけ最近変更されたかをランク付けの基準にすることで精度が上がることがわかる

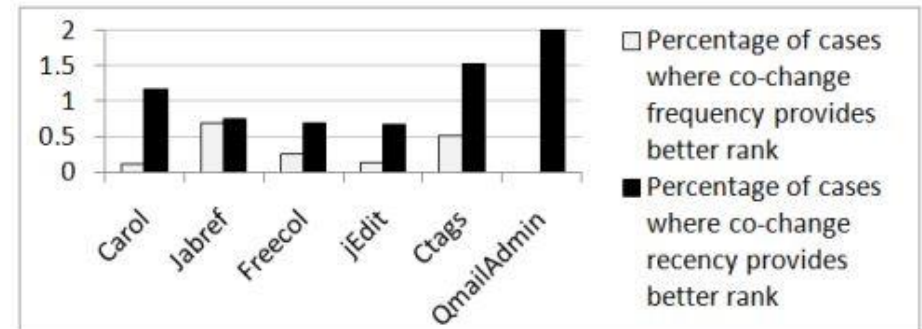


Figure 8: Comparison of the two ranking systems (ranking using co-change frequency and co-change recency)

RQ3で何をするか

- RQ1のFig4でEVを示さないクローン片が大多数であることがわかる
- 大多数のEVを示さないクローン片でもランク付けを行いたい
- EVを示すクローン片から特徴を見つけ出し, EVを示さないクローン片でも同じようにその特徴を持っているかどうかを調べる
- その特徴を持っているなら, その特徴を用いてランク付けをすることが出来ると考えた

EVを持つクローン片の特徴

- 次の特徴について分析する
 - ◆ クローン片同士がどれだけ近い
か
 - ◆ クローン片がメソッドかそれ以外
のブロックか

- 右図はクローン片同士が同じ
ファイルにある割合とクローン
片がメソッドである割合を示して
いる

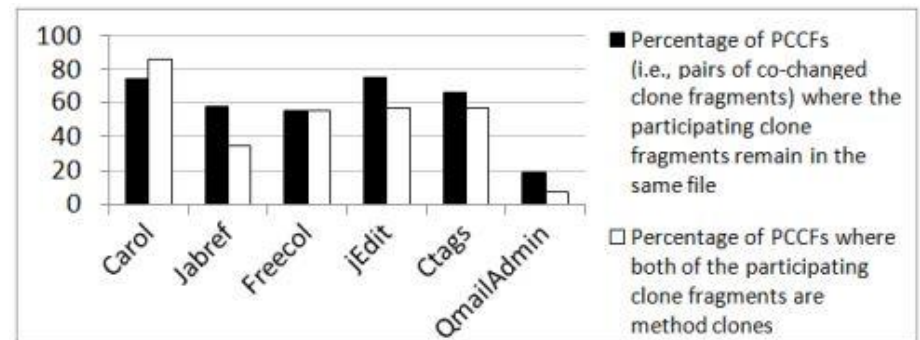


Figure 9: Comparison of the characteristics of the clone fragments that exhibit evolutionary coupling

EVを持たないクローン片の特徴

- 右図はEVを持たないクローン片に対してクローン片同士の近さをランク付けの基準にした方が精度がよかった割合とクローン片がメソッドであることを基準にした方が精度がよかった割合を示したもの

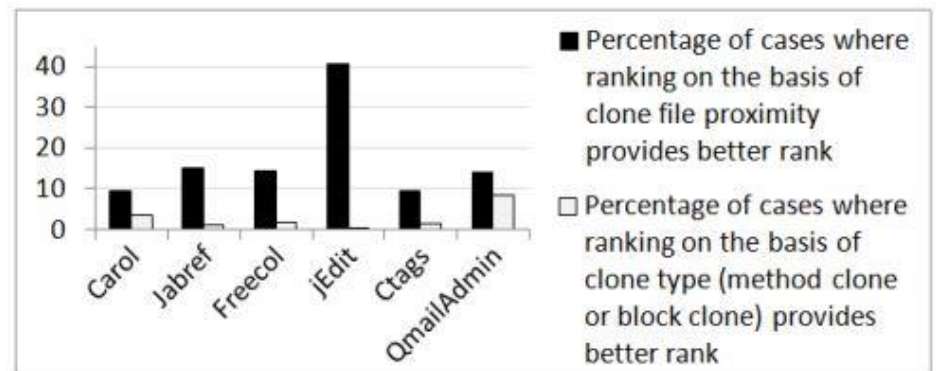


Figure 10: Comparison of the two ranking systems for the non-predicted co-change candidates

- クローン片同士の近さをランク付けの基準とした方が精度がよいことがわかる

RQ4で何をするか

- RQ3で何をランク付けの基準とすることでよりよい精度が得られることがわかったところで、EVを持つクローン片とEVを持たないクローン片の両方に対してランク付けを行いたい
- どれだけ最近変更されたかとクローン片同士の近さを組み合わせてランク付けの基準とする場合とクローン片同士のみをランク付けの基準とする場合のどちらが精度がよいかを調べる

ランク付けの精度

- 右図は組み合わせた場合の精度の良さとクローン片同士の近さのみを基準とした場合の精度の良さを示したものである
- 組み合わせた場合のほうが精度がよいことがわかる

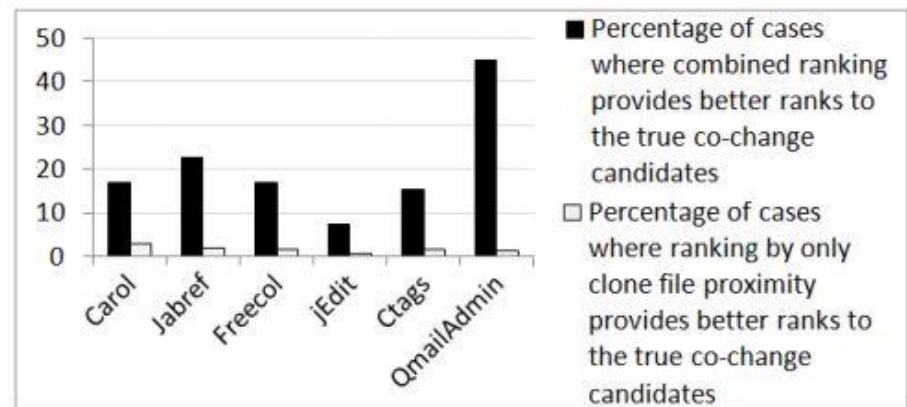


Figure 11: Comparison between the combined ranking (i.e, ranking predicted candidates by co-change recency and non-predicted co-change candidates by clone file proximity) and ranking considering only clone file proximity

まとめ, 所感

まとめ

- EVを用いることでかなり良い適合率で共変化するクローンの予測が出来ることがわかった
- EVを持つクローン片のランク付けはどれだけ最近変更されたかをランク付けの基準とすることで精度が上がることがわかった
- EVを持つクローン片と持たないクローン片の両方のランク付けはどれだけ最近変更されたかとクローン片同士の近さを組み合わせた基準とすることである程度良い精度でランク付けすることが出来ることがわかった

所感

- 共変化するクローン片の予測だけなら現場でも使えそうだった
- ランク付けの精度をランクの和の小ささで決めてもいいのか疑問に思った
- EVを持つクローン片の特長をなぜあれに決めたのか書いていなかったのを書いてほしい