

検出済みクローンペアの推移的關係を利用した 未検出クローンペアの補完手法

2026/07/23

○石村 涼介, 肥後 芳樹 (大阪大学)



コードクローン

互いに一致・類似した部分を持つソースコード片^[1]

クローン関係にあるコード片のペアを**クローンペア**と呼ぶ

クローンはバグを伝搬し^[2]システムの保守性を損なう

クローン検出は重要

多くの**コードクローン検出ツール**が提案されている

[1] 井上克郎, 神谷年洋, 楠本真二, コードクローン検出法, コンピュータソフトウェア, Vol. 18, No. 5, pp. 529–536, 2001.

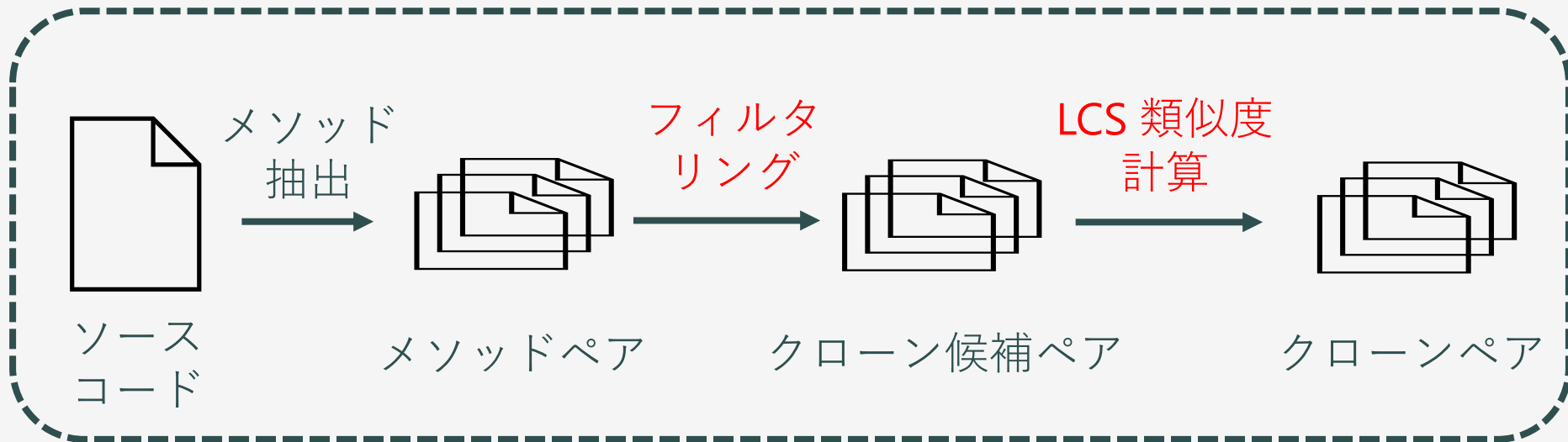
[2] M. Mondal, C. Roy, and K. Schneider, “A Fine-Grained Analysis on the Inconsistent Changes in Code Clones,” In 2020 IEEE ICSME, pp. 220–231, 2020.

メソッド単位のクローンペアを検出するツール

LCS 類似度がしきい値以上のペアをクローンペアとする

LCS : Longest Common Subsequence (最長共通部分列)

LCS の計算は重いため前処理として**フィルタリング**する



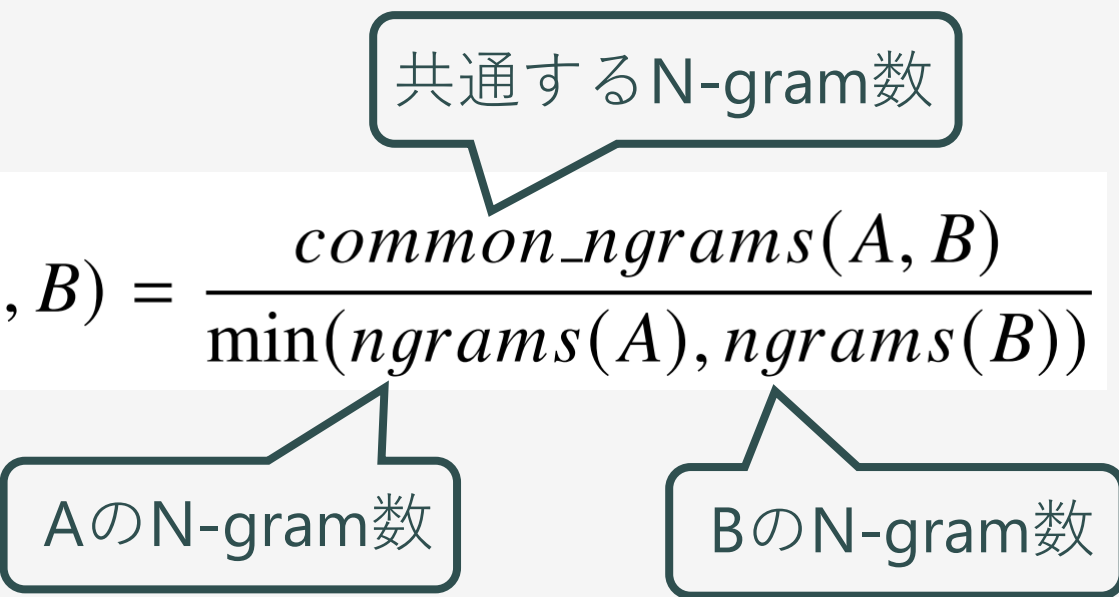
NIL のフィルタリング (N-gram類似度)

メソッドA, Bのソースコードに対して以下を計算

しきい値を下回るペアは本処理に進まない

99%以上のペアがフィルタリングで除外される

検出すべきクローンが誤って除外される可能性がある

$$filtration_sim(A, B) = \frac{\text{共通するN-gram数}}{\min(\text{AのN-gram数}, \text{BのN-gram数})}$$


NIL の本処理 (LCS 類似度)

メソッドA, Bのソースコードに対して以下を計算
しきい値以上のペアをクローンペアとして出力
フィルタリングで除外されたペアは計算されない

$$verification_sim(A, B) = \frac{lcs(A, B)}{\min(|A|, |B|)}$$

LCSのトークン数

Aのトークン数

Bのトークン数

フィルタリングと検出性能のトレードオフ

フィルタリング条件が**厳しい**場合

パスするペアが**少ない**

○ 本処理の実行時間が**短い**

× 見逃しが**多い**

フィルタリング条件が**緩い**場合

パスするペアが**多い**

○ 見逃しが**少ない**

× 本処理の実行時間が**長い**

本研究の概要

目的

フィルタリングと検出性能の**トレードオフの緩和**
(**見逃しが少ないかつ実行時間が短い**クローン検出)

方法

クローン関係の推移律を用いた後処理を提案

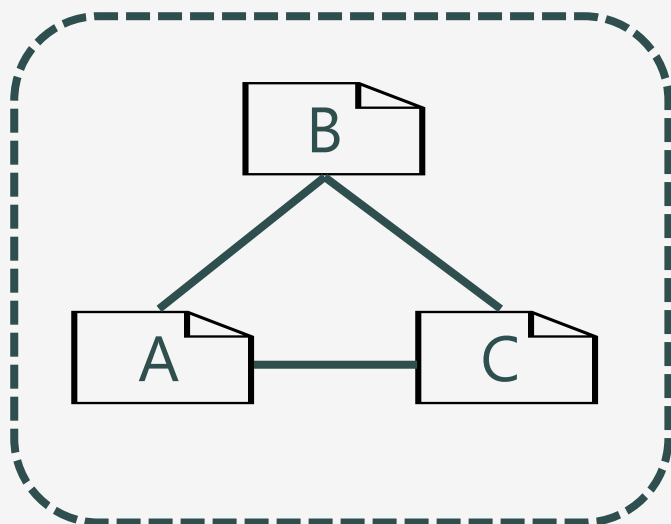
クローン関係の推移律

クローン関係の推移律は**一般的には成立しない**

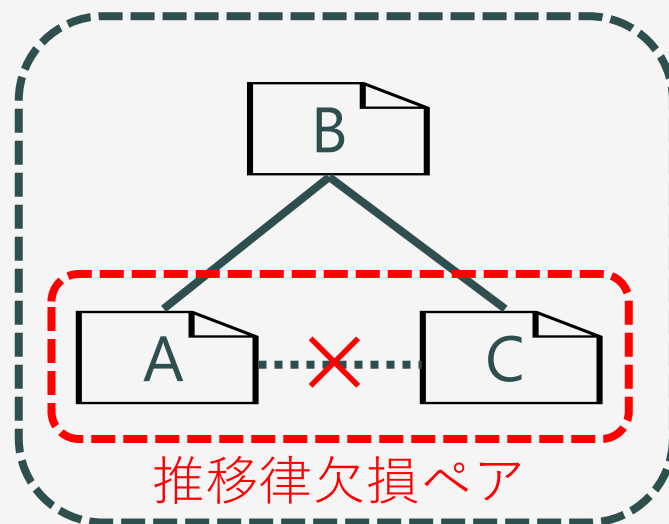
本研究では**推移律欠損ペア**を定義

メソッドAとB, BとCがそれぞれクローンのとき,

AとCがクローンでない場合, **AとCを推移律欠損ペア**と呼ぶ



推移律が成立する例



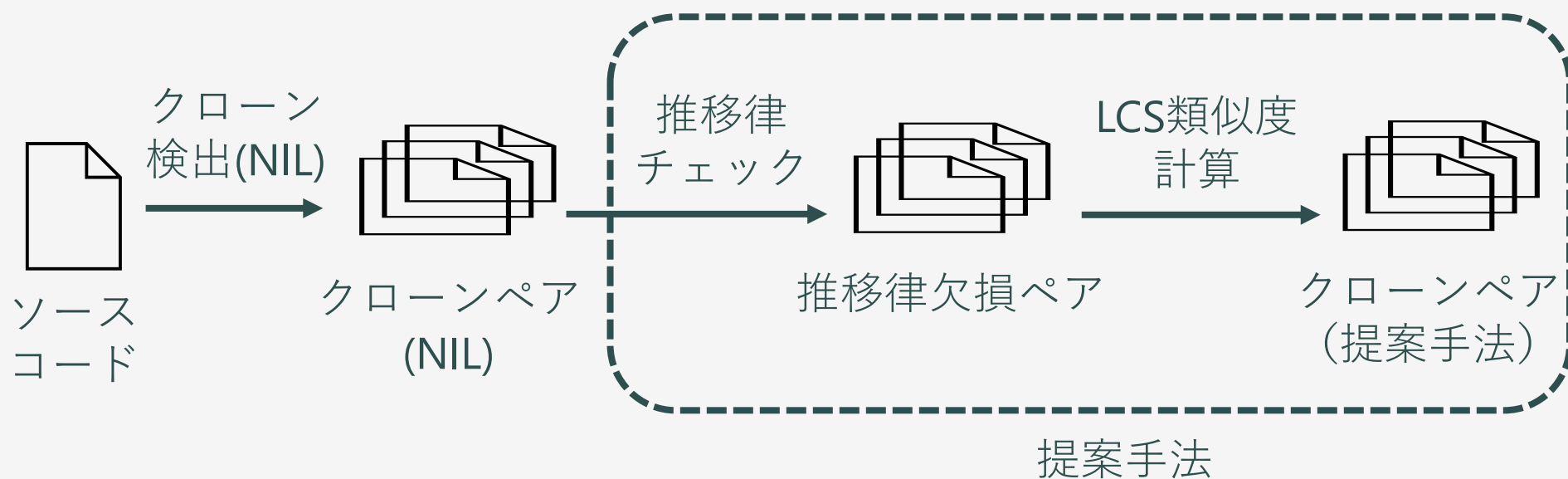
推移律が成立しない例

提案手法

推移律欠損ペアに対してLCS類似度を計算する

[仮説]推移律欠損ペアはクローンペアの可能性が高い
フィルタリングで除外されたペアを**補完**

NILの後処理



RQ (Research Question)

2つのRQを設定

RQ1：NILの見逃しをどの程度補完できるか

RQ2：実行時間のオーバーヘッドはどの程度か

評価指標

RQ1：Recall

フィルタリングを一切実行しない場合の検出結果に対するRecall

RQ2：実行時間

対象ソフトウェア

Java OSS を使用(テストコードは除く)

NILにおいてフィルタリングを
一切実行しない場合の値



ソフトウェア	LOC	メソッドペア数	クローンペア数	検出時間[s]
Maven 3.9.16	55,978	1,175,811	7,217	2.5
Ant 1.10.17	114,196	3,918,600	1,435	5.1
Tomcat 10.1.55	266,124	21,968,506	17,775	17.8
Lucene 10.4.0	491,545	71,862,066	39,067	43.5

Recallの基準集合

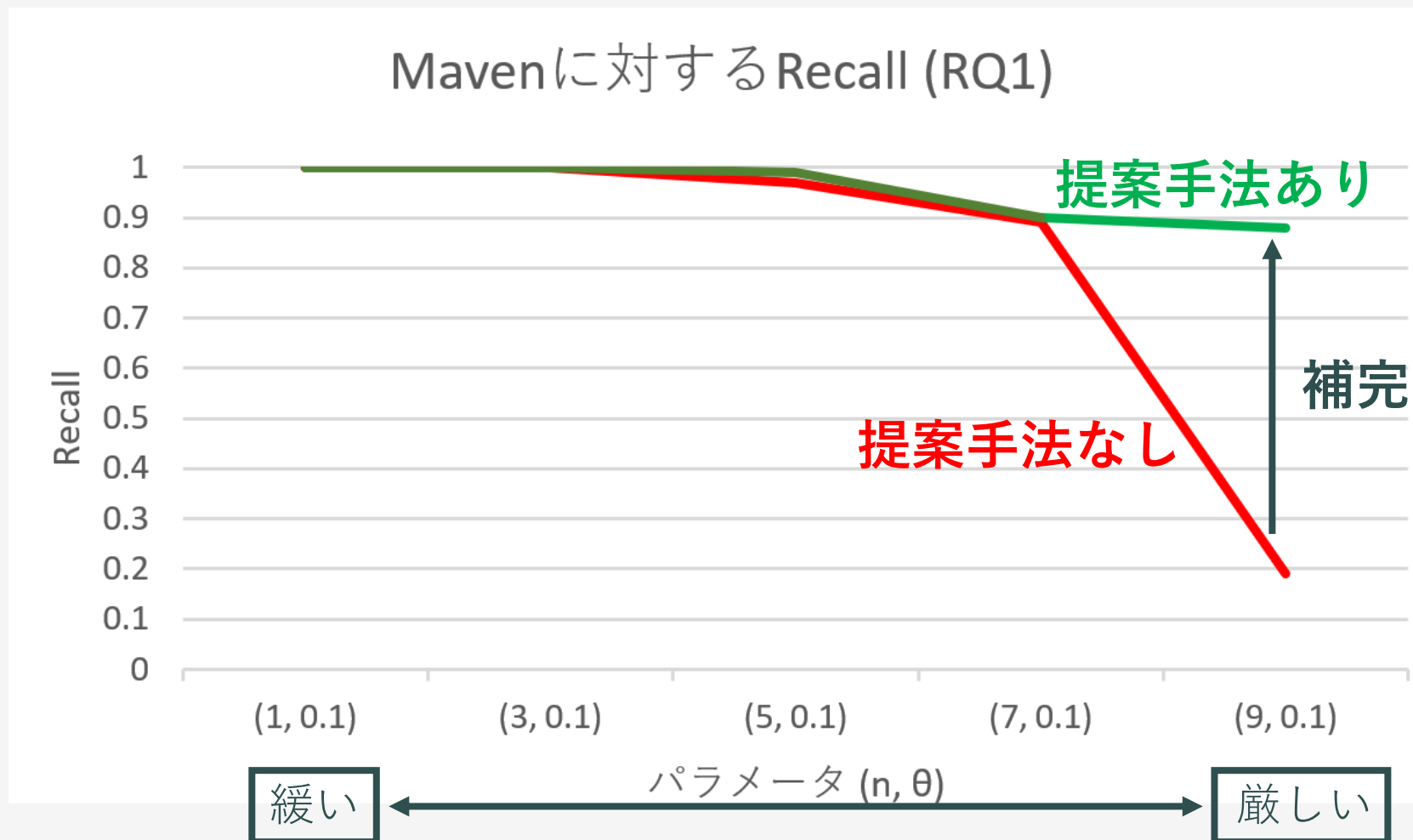
パラメータ設定

パラメータの値が大きいほどフィルタリングが厳しい

パラメータ	意味	デフォルト値	設定値
n	N-gram の長さ	5	1, 3, 5, 7, 9
θ	フィルタリングのしきい値	0.1	0, 0.1, 0.2, ... , 1.0
δ	LCS 類似度のしきい値	0.7	0.7(固定)

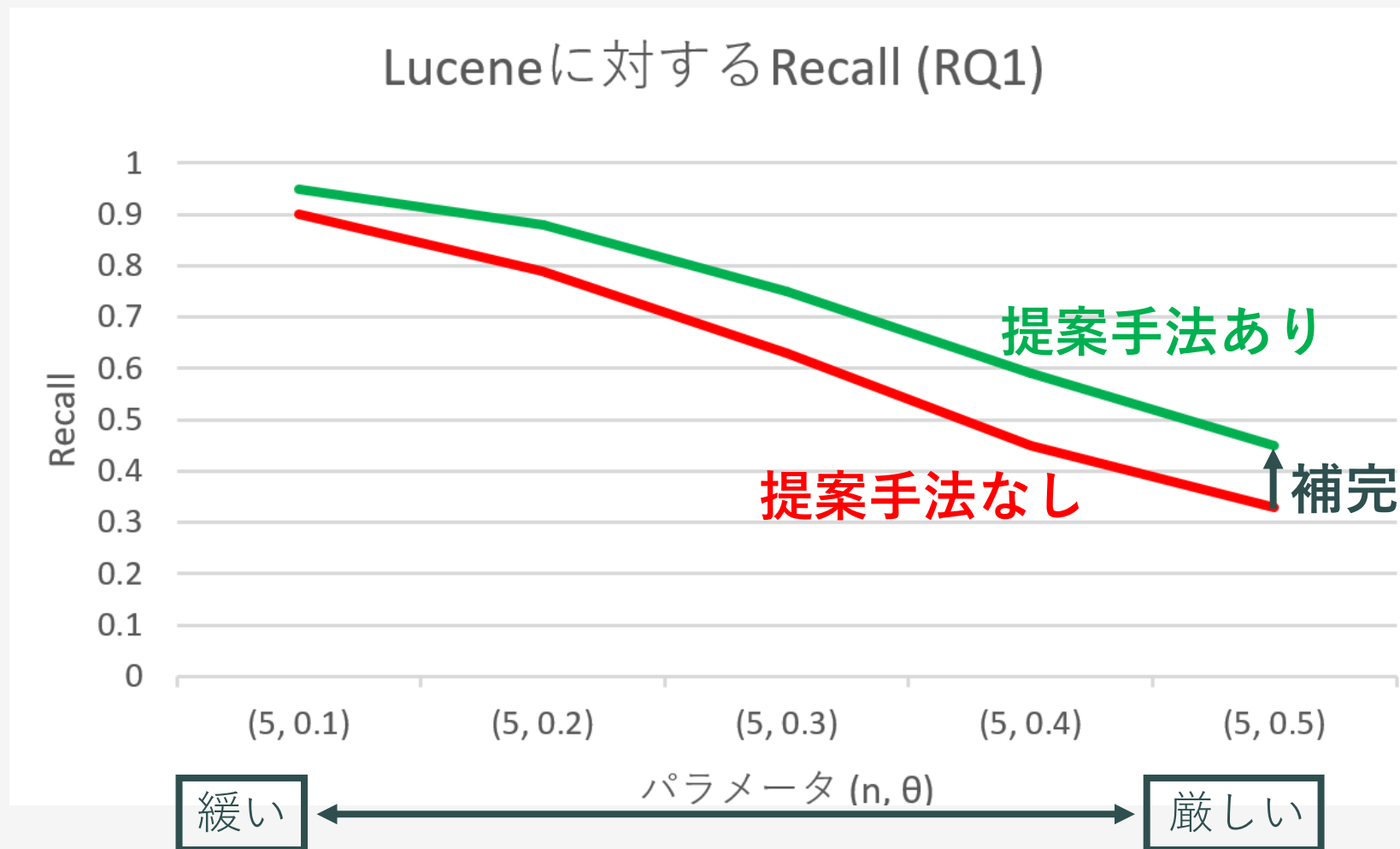
RQ1：見逃しの補完性能(パラメータn)

提案手法はRecallの低下を緩和する



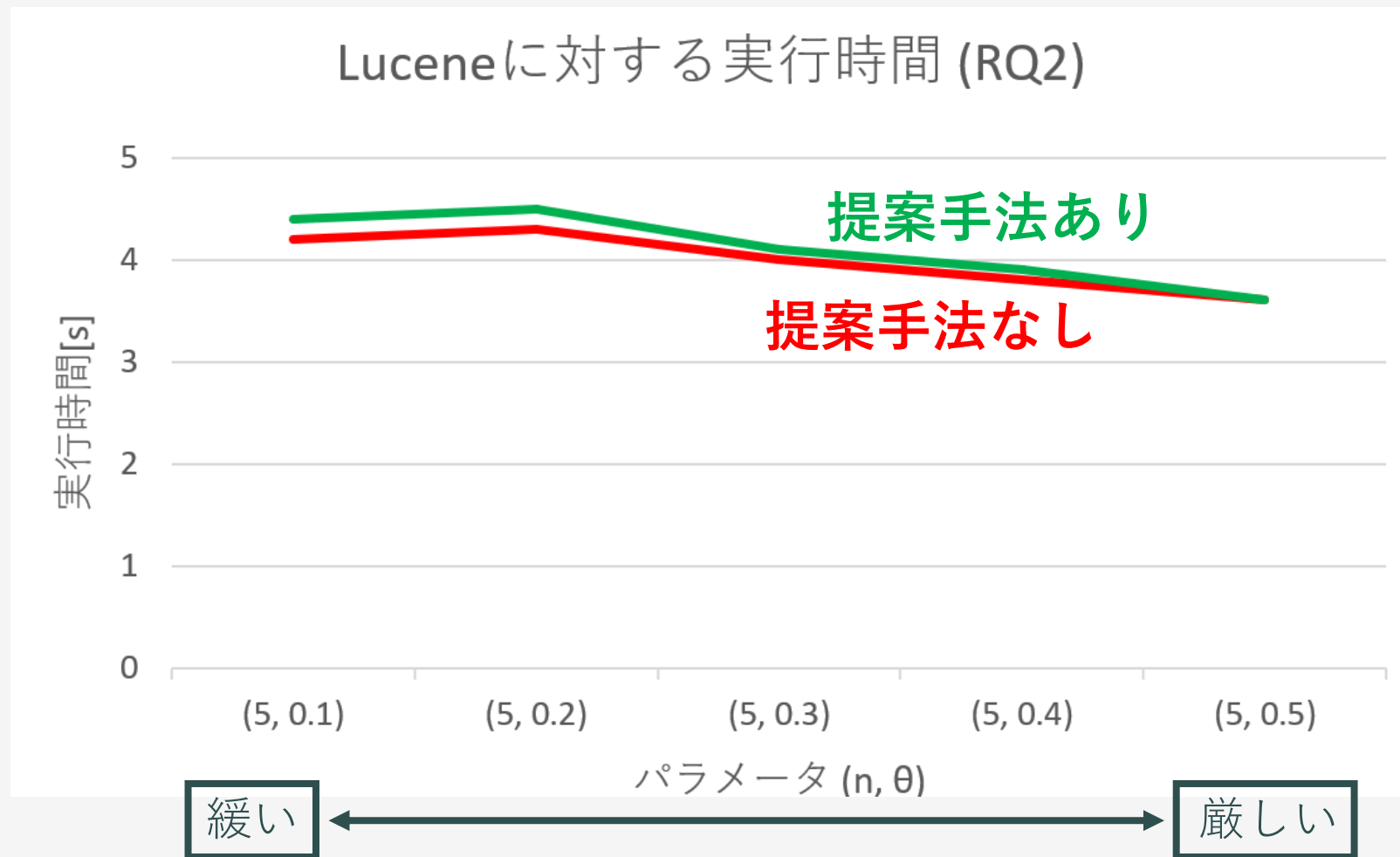
RQ1：見逃しの補完性能(パラメータ θ)

提案手法はRecallの低下を緩和する



RQ2：実行時間のオーバーヘッド

提案手法のオーバーヘッドは小さい (最大10%程度)



まとめ

NILの**後処理としてのクローンペア補完手法**を提案

推移律欠損ペアに注目することでRecallを向上

実行時間のオーバーヘッドは**全体の10%程度**

今後の展望

NIL以外のクローン検出ツールに対して適用