

nokogiri(仮)

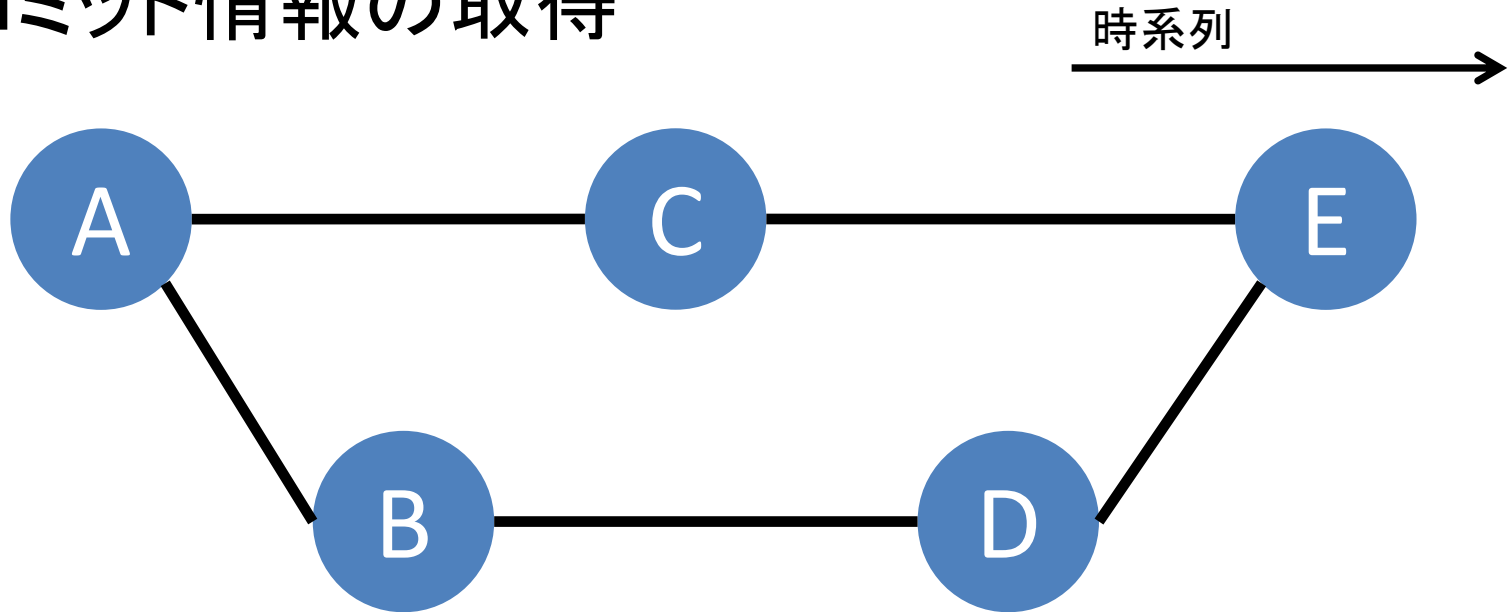
コードクローン追跡方法

動作手順

- STEP1: コミット情報取得
- STEP2: 最初のコミットにおける, ファイル, コード片, クローンセットの特定処理
- STEP3: 次のコミット(最初のコミット以降)における, ファイル, コード片, クローンセットの特定処理
- STEP4: クローンセットの前後関係を特定

STEP1

- コミット情報の取得



- コミット情報

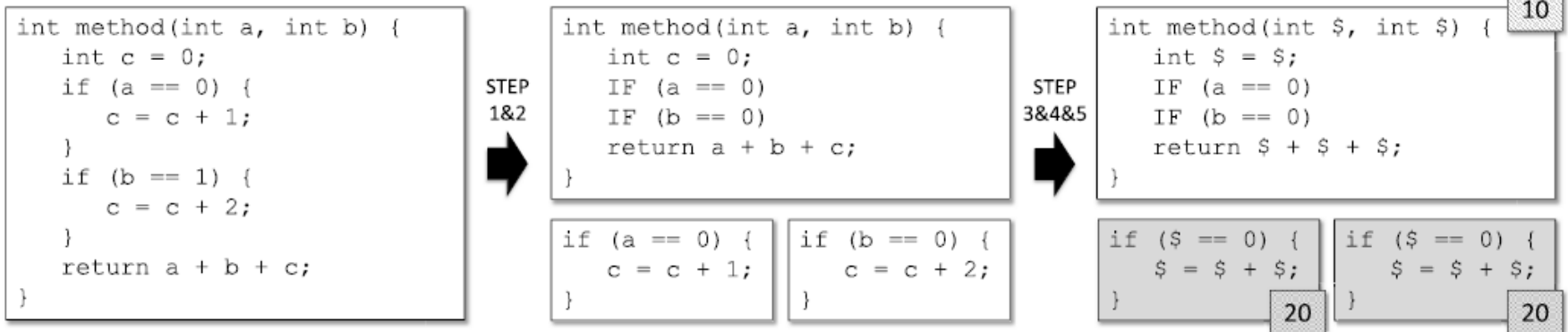
- コミット日時
- コミッタ名
- メールアドレス

- コミットの親子関係

- A – 親: null, 子: C, D
- B – 親: A 子: D
- C – 親: A 子: E
- D – 親: B 子: E
- E – 親: C, D 子: null

STEP2

- 最初のコミットにおけるファイル, コード片, クローンセットの特定
 - 全てのファイルに対してスキャンを行い, コード片クローンセットを特定



ブロックの特定

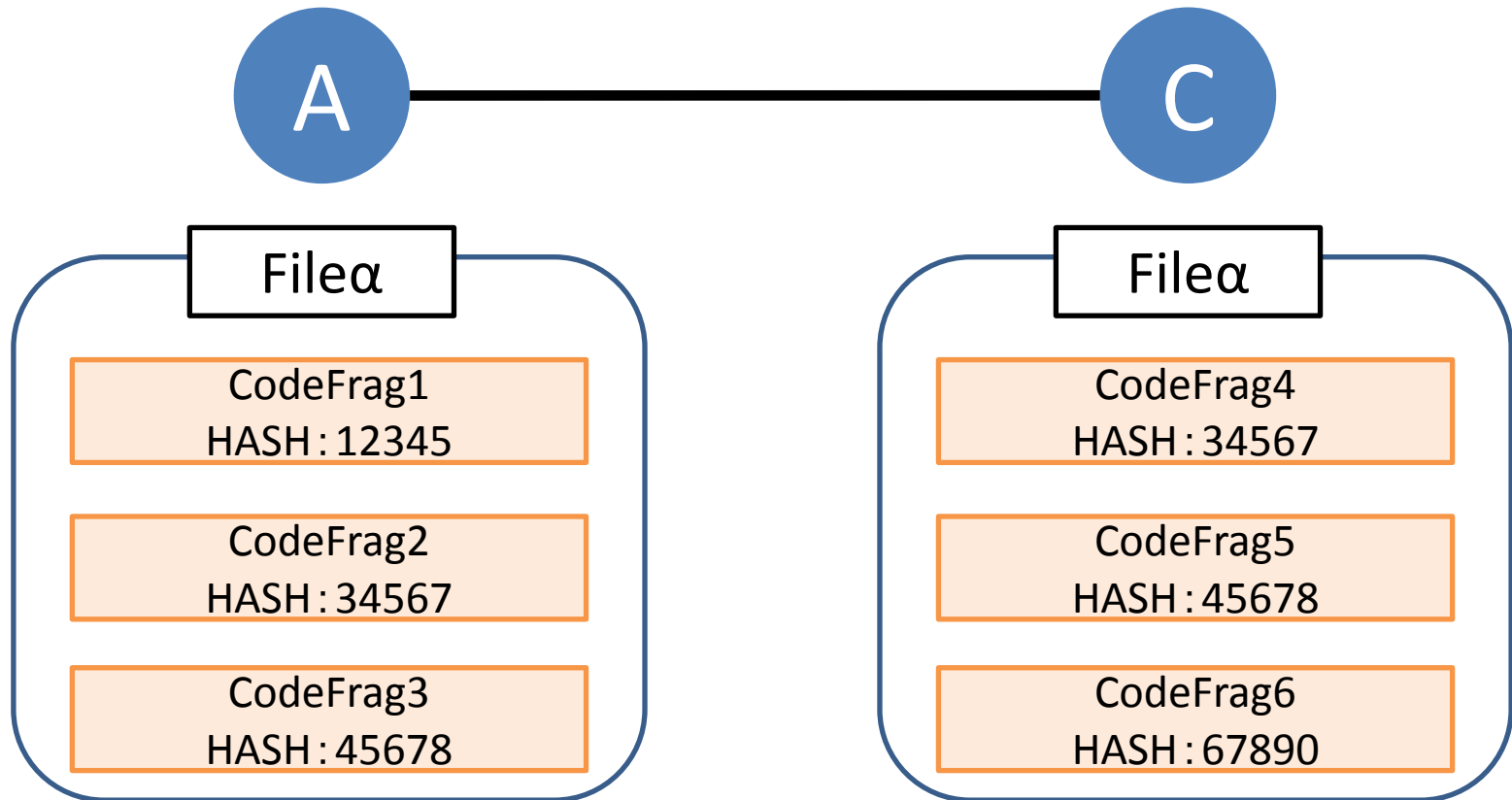
正規化
ハッシュ値の算出
コードクローン特定

STEP3

- 次のコミット（最初のコミット以降）における、ファイル、コード片、クローンセットの特定
 - STEP2と同じようにすべてのファイルをスキャンする方法では時間がかかりすぎる（特にコード片特定処理にかかる時間が長い）
 - 前のコミットに存在するコード片（1）
 - 削除されたコード片（2）
 - 新規追加されたコード片（3）
 - 次のコミットにおけるコード片 = $(1) - (2) + (3)$

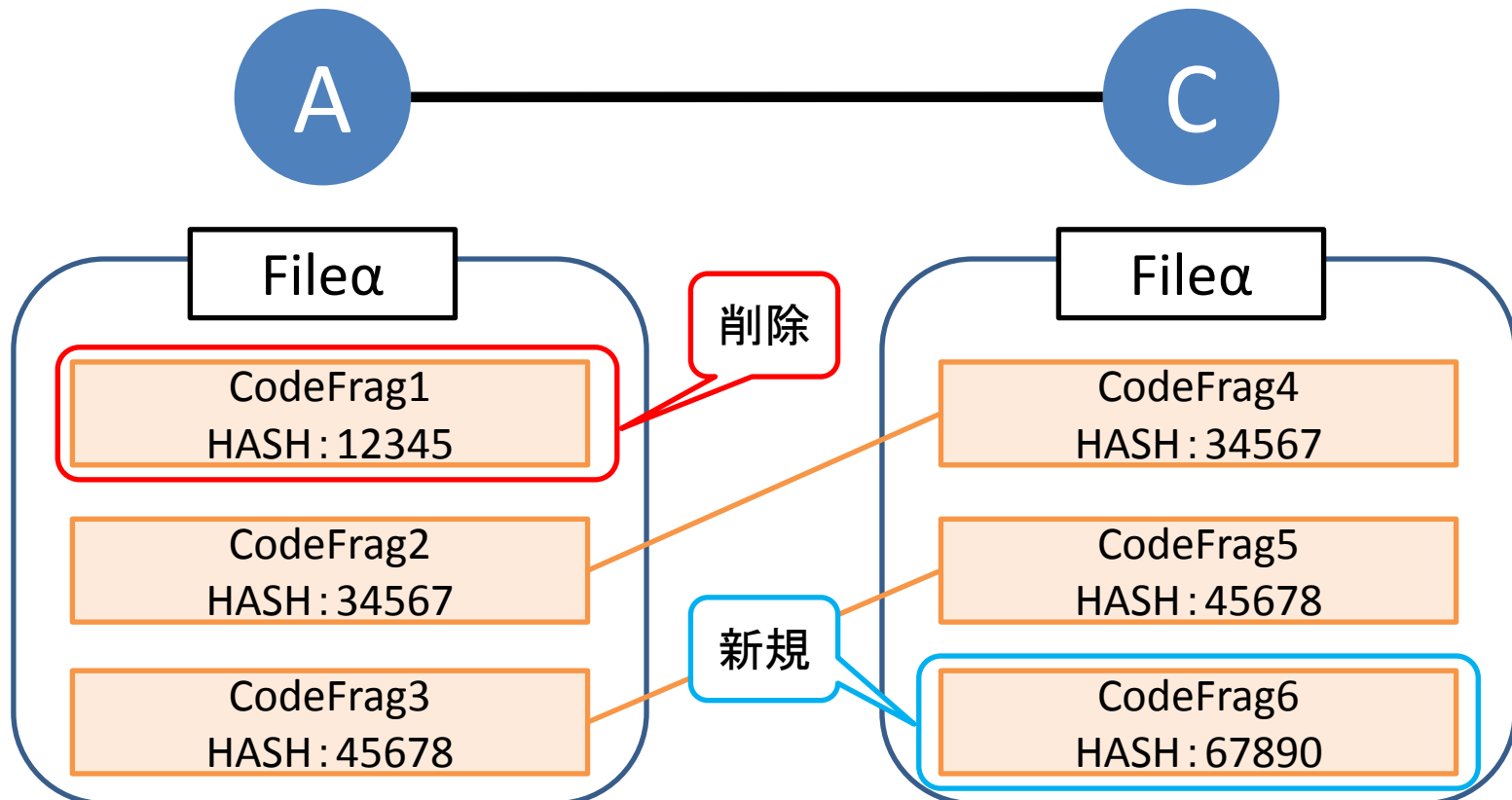
STEP3-1

- ひとつ前のコミットとの差分を取り，変更があったファイル情報を取得し，コード片特定



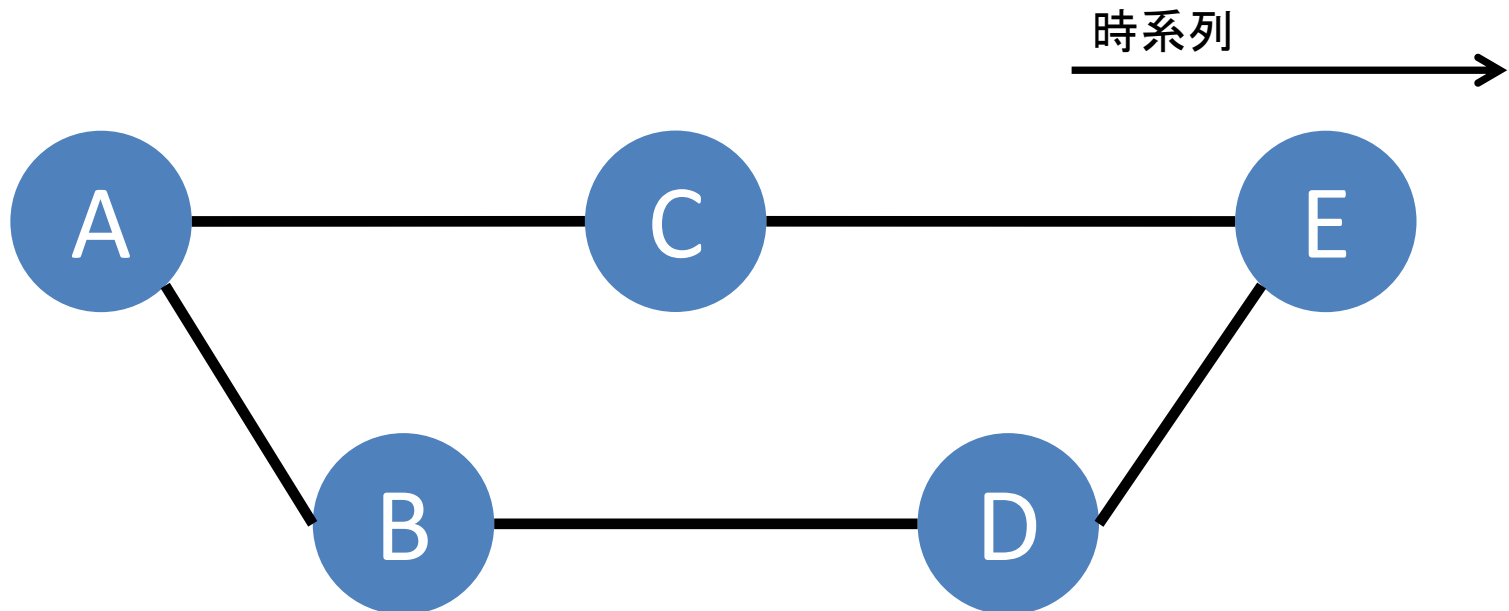
STEP3-2

- ソースコード文字列をもとに算出されるHASH値が
 - Aにのみ存在するコード片は削除されたコード片
 - Cにのみ存在するコード片は新規追加されたコード片



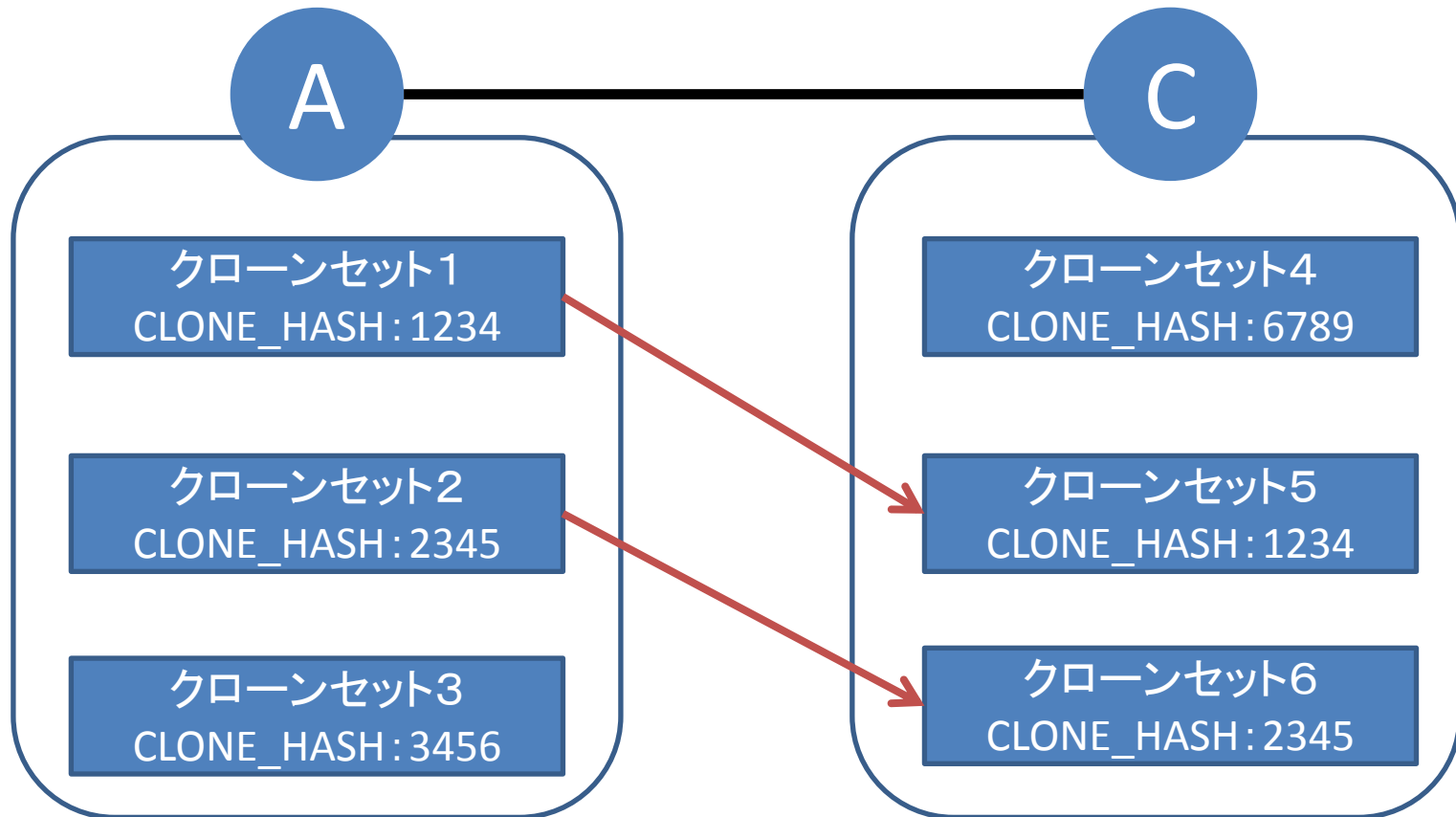
STEP3-3

- STEP3-1, STEP3-2を繰り返す
 - 各コミットに存在するコード片, クローンセット情報をDBに格納する



STEP4

- クローンセットの前後関係を特定
 - CLONE_HASH(正規化した後のHASH値)が一致するものに対してリンクをはる



参考

- apache struts
 - コミット(2015年3月1日)
 - f7364a50dc6a82b51a57dab9b3efd6303e2295b6
 - ファイル数: 1955
 - コード片特定処理時間(スレッド数4): 53738[ms]
 - コード片数: 14950
 - クローン特定処理時間: 2483[ms]