クラスタリングを用いた ソフトウェアパターン分類支援手法

久保淳人† 鷲崎弘宜‡ 深澤良彰†

(†早稲田大学 ‡国立情報学研究所)

ウィンターワークショップ・イン・道後2008

2008年1月24日

背景

- ソフトウェアパターン：ソフトウェア開発における特定の文脈で繰り返し発生する問題と、問題に対する実証済の解法、解法を導く制約条件などの記述

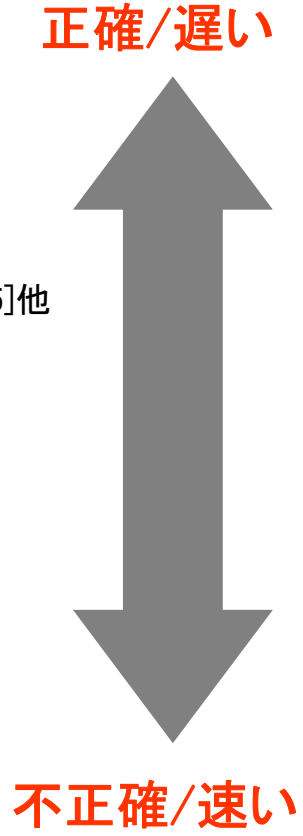
- ソフトウェアパターンの増加：約2000個（2007年9月）[Henninger2007]
  - PLoP系会議でレビュー、もしくは、出版済

- ソフトウェアパターンの利用プロセス[Washizaki2003]

  選択→拡張→利用→評価
  - 選択は重要：後のプロセスすべてに影響
  - 状況に適したパターンを選ぶためには、各パターンの内容を把握する必要がある
    - 困難：(多数のパターンからランダムに選べば)ほとんどのパターンは現在の状況にマッチしない

課題と現在までのアプローチ

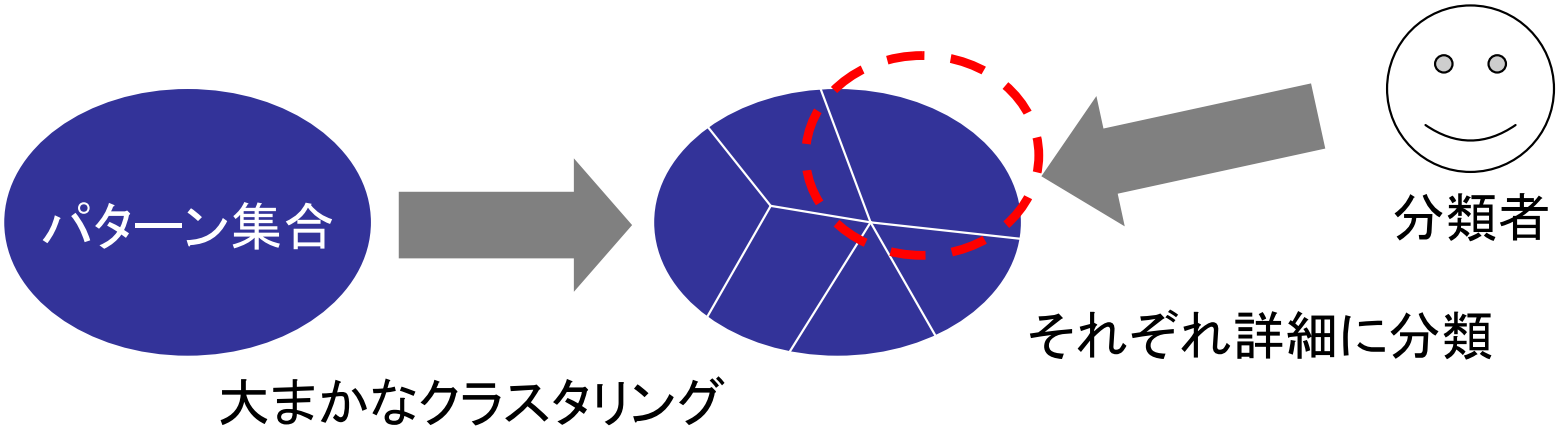
- 問題点：多数のパターンから、現在の状況に適したパターンを選択するのは難しい
  - →候補を絞る
  - →パターン分類
- 現在までのアプローチ
  - 手動：
    - カテゴリ分類（例: 適用対象/目的 [Gamma1994]他）
    - パターンマップ：パターン間の関係性を示す [Zimmer1995]他
  - 半自動：
    - 手動入力の各評価値の類似性をバネモデルで表示 [Frauenberger2007]
    - 手動入力の各評価値を指定して検索[Vrsalovic]
  - 自動：
    - パターン間関連の自動分析 [Kubo2005]



課題の検討・提案手法

- 課題:
  - 自然言語処理による詳細な自動分類は精度に難あり
  - 多数のパターンの手動管理は手間がかかる

- 提案手法：文書クラスタリングを用いたソフトウェアパターン分類
  - 自動・大まかな分類
  - 手動・詳細な分類

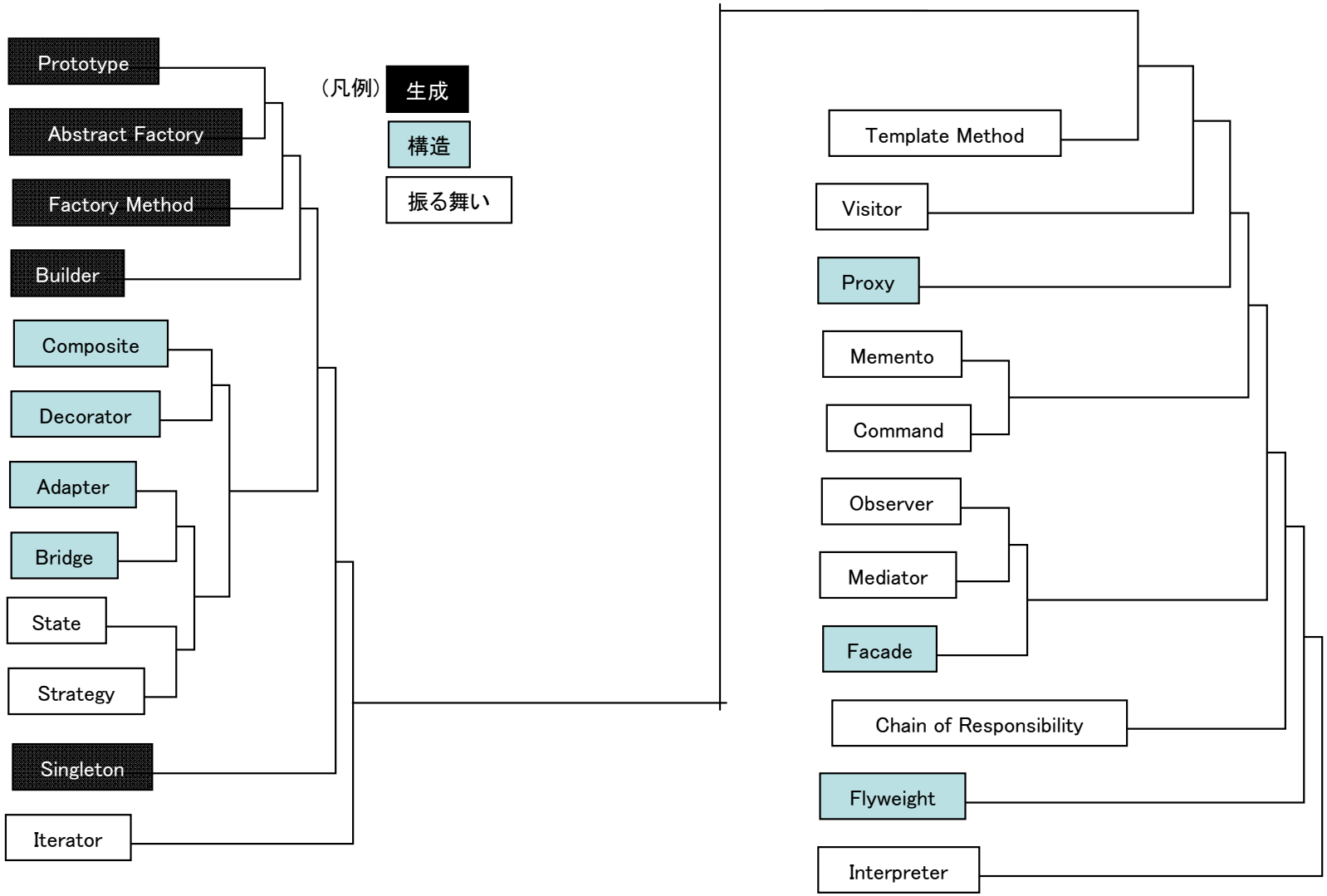


クラスタリング手法

- 階層型
  - 凝集型: 初期状態として1要素のクラスタを作成、最も近いクラスタを併合する、という操作を繰り返す
    - クラスタ間距離の算出法でバリエーション
    - 予備実験では、クラスタ内文書を単純連結した文書を新クラスタの代表文書として距離を算出
  - 分岐型 (あまり使われない)
- 非階層型
  - k-mean法: 評価関数を定義し、ランダムに作成したk個のクラスタ間で、評価関数が最適になるように要素を入れ替え

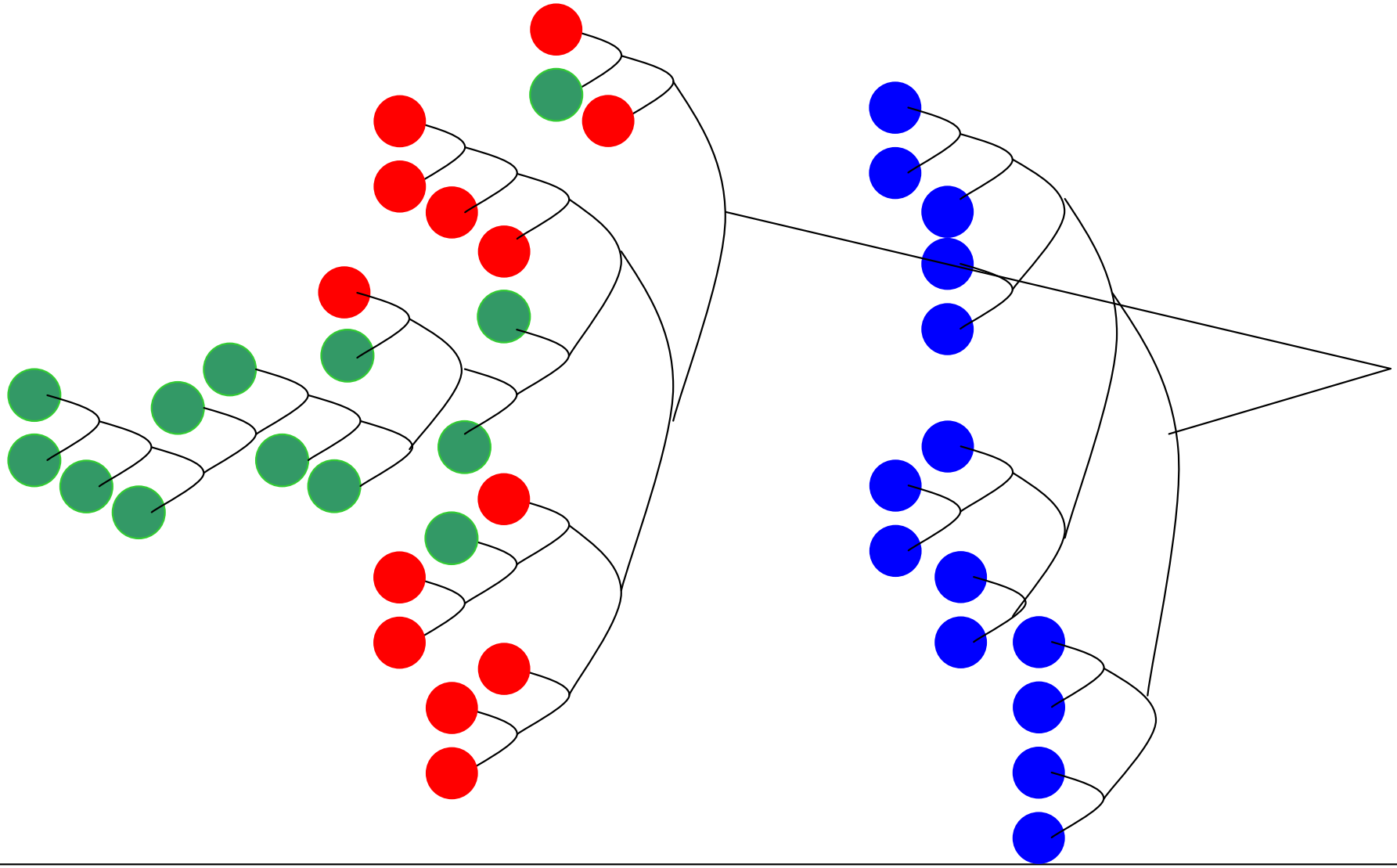
予備実験 (1)

- GoFデザインパターンについて、階層/凝集型クラスタリングを実施し、デンドログラム(樹形図)を描画



予備実験 (2)

- GoF/PoSA/Organizationalパターンについて、同様のクラスタリングを実施 (図は一部)



まとめ・議論

- 提案手法：文書クラスタリングを用いたソフトウェアパターン分類
  - 自動分類は精度に難あり
  - 人手の分類は大量のパターンを扱えない

  →大まかに自動分類、細かい分類は人手で

- 議論
  - パターン利用支援にパターン分類は有効か
  - 大規模なパターン集合についての、手動での分類は現実的か
  - 一部のパターン分類を自動化するとして，手動でのパターン分類との境界はどの程度の規模にするのが妥当か
  - クラスタリング手法の選択