ICDE2016 & WWW2016 勉強会

天笠俊之

筑波大学

amagasa@cs.tsukuba.ac.jp

5C: Clustering

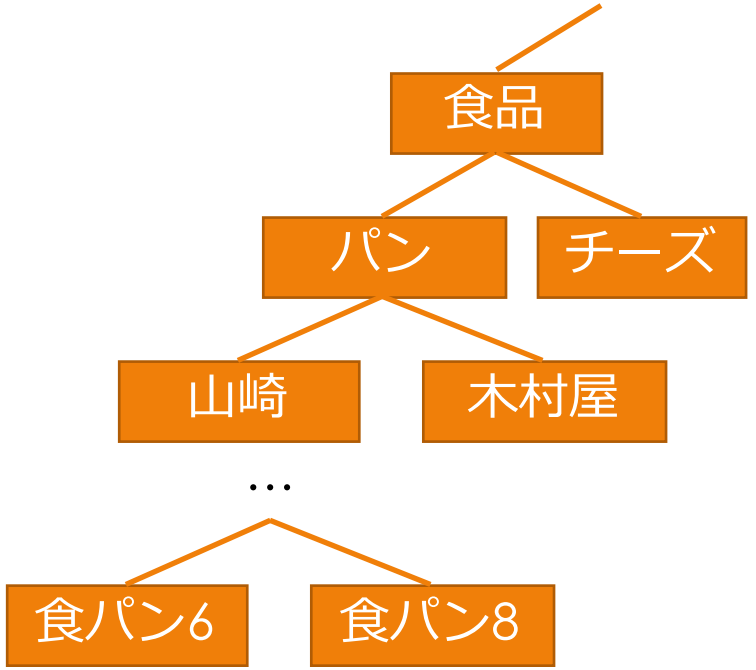
PurTreeClust: A Purchase Tree Clustering Algorithm for Large-scale Customer Transaction Data

Xiaojun Chen (Shenzhen University), Zhexue Huang (Shenzhen University), Jun Luo (SIAT)

研究の背景

- やりたいこと
  - 購買履歴トランザクションに対するクラスタリング
  - 類似した購買行動をとるユーザをグループ化
- 問題
  - 商品点数は膨大（>10k）
  - カテゴリ階層も複雑（>1k）
  - 商品IDによる単純な集約ではうまくいかない

TID	Items
1	牛乳，ビール
2	パン，牛乳，ビール
3	ビール，おむつ
4	チーズ

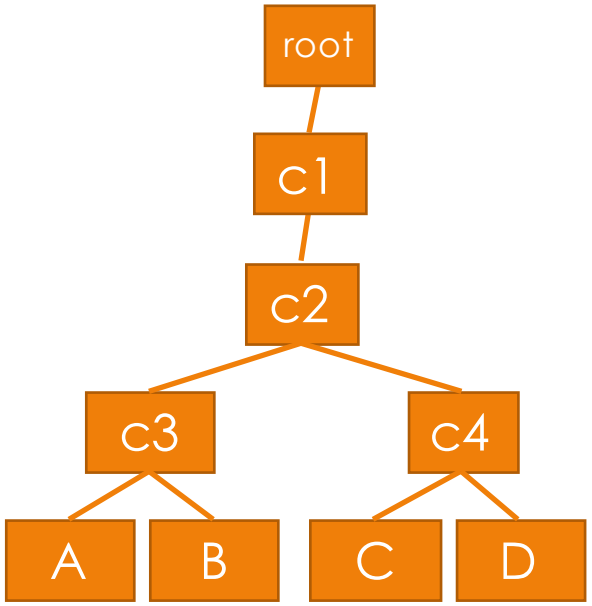


本研究の目的・貢献

- 目的
  - 購買履歴と製品カテゴリから導出されるデータ構造（Purchase Tree）のためのクラスタリングアルゴリズムを提案.
- 貢献
  - Purchase treeのための新たな距離（PurTree距離）を提案.
  - 大規模な購買履歴データを効率良く扱うため，cover tree を利用した索引手法と，初期クラスタを発見するための密度推定手法を提案.
  - クラスタリングアルゴリズム PurTreeClust を提案. 計算量は $O(c^6 + k)\, n\, log\, n)$. ただし，n は purchase tree の数，k はクラス多数，c は拡張係数.
  - クラス多数の推定手法も提案.

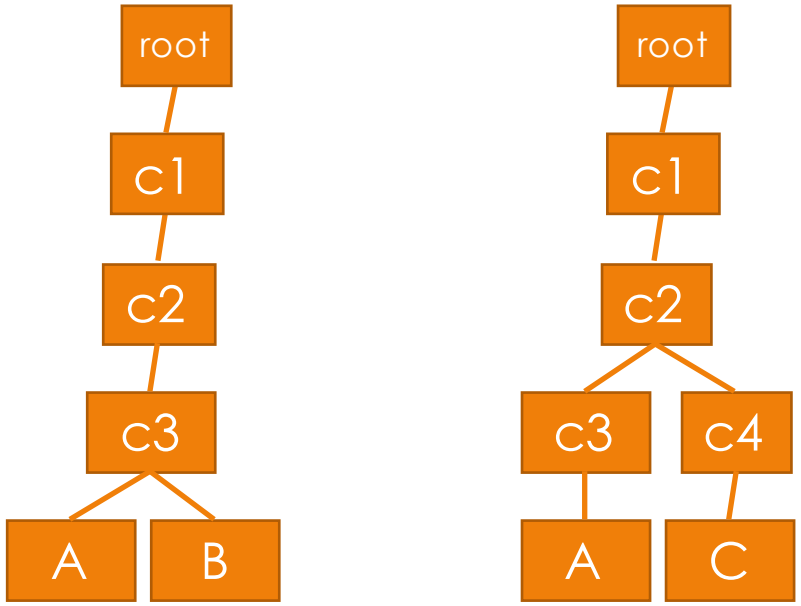
Product tree 及び Purchase tree

製品カテゴリ（product tree）



仮想的な根を持つ．全ての葉は同じレベルにあることを仮定

Purchase tree

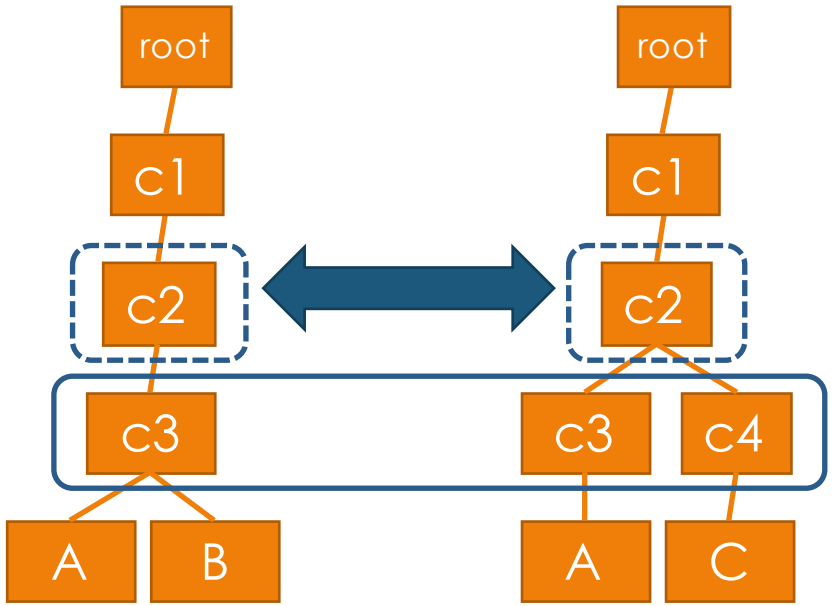


根から購入した製品（葉）までのパスから構成されるproduct treeの部分木

PurTree距離

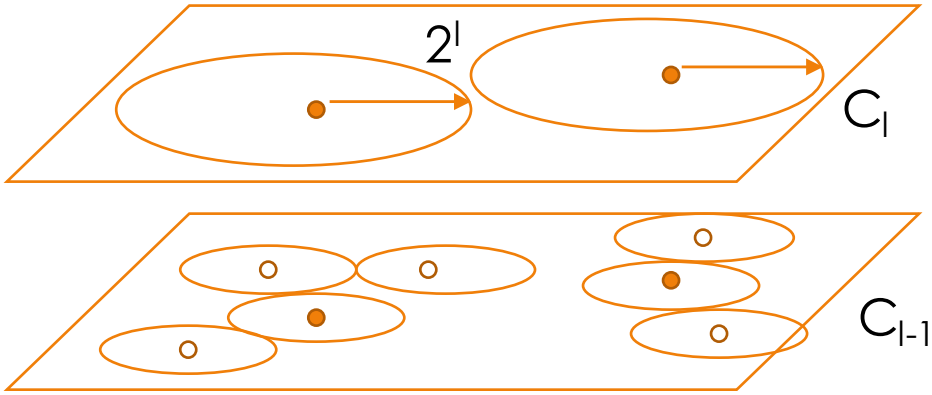
- レベルlにおけるproduct tree間の距離
  - 各カテゴリの子ノードの一致率（Jaccard距離）の重み付き和
- PurTree距離
  - 各レベルでの距離の重み付き和
  - パラメータγにより，パスの影響を調整可能
- PurTree距離が距離の公理を満たすことを証明

例：レベル3での距離



Cover treeの利用

- Cover tree [Beygelzimer et al., 2006]
  - 距離空間のための木構造索引．最近傍検索のために使われる．
- Purchase treeへの適用
  - Cover treeの根から順に投入．
- オブジェクトpの密度推定
  - Pからの距離が 2^l 未満のオブジェクトをCover tree を使って探索



Cover treeのイメージ

PurTreeClustアルゴリズム

- PurTreeClustアルゴリズム
  1. Cover tree CT で，k 以上のオブジェクトが存在するレイヤーを探索
  2. 各オブジェクト周辺の密度を計算
  3. 密度の高いオブジェクトを初期セントロイドとして各オブジェクトをクラスタに割当
- クラスタ数の推定
  - ギャップ統計を利用

実験

- データセット
  - 中国のスーパーマーケットの購買履歴データ
    - 顧客数 >40k
    - Product treeの製品数 22,666
    - ランダムサンプリングにより10件のデータセットを生成
  - 比較手法
    - 階層型クラスタリング（HAC），スペクトラルクラスタリング（SPEC），DBSCANと比較

実行時間／クラスタリングの結果

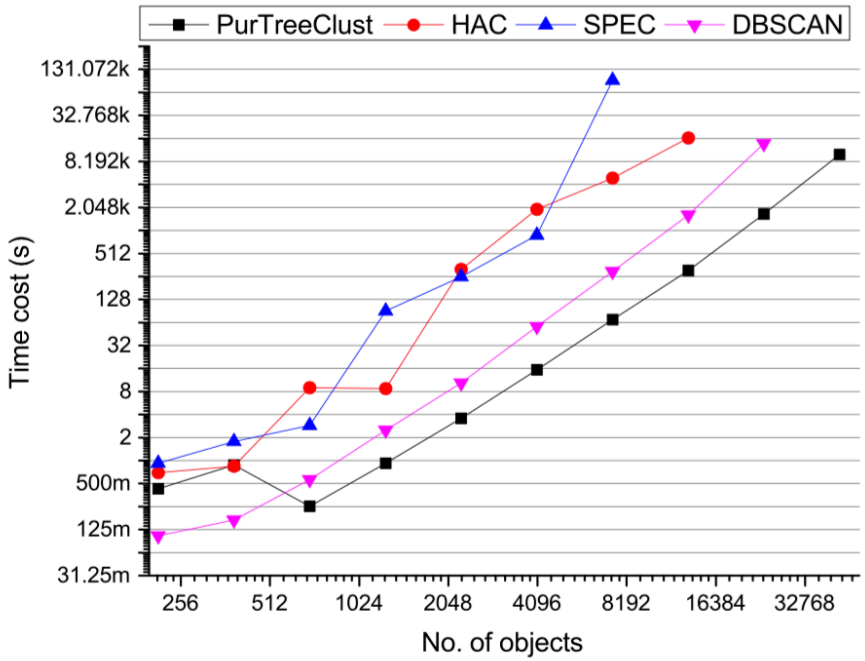
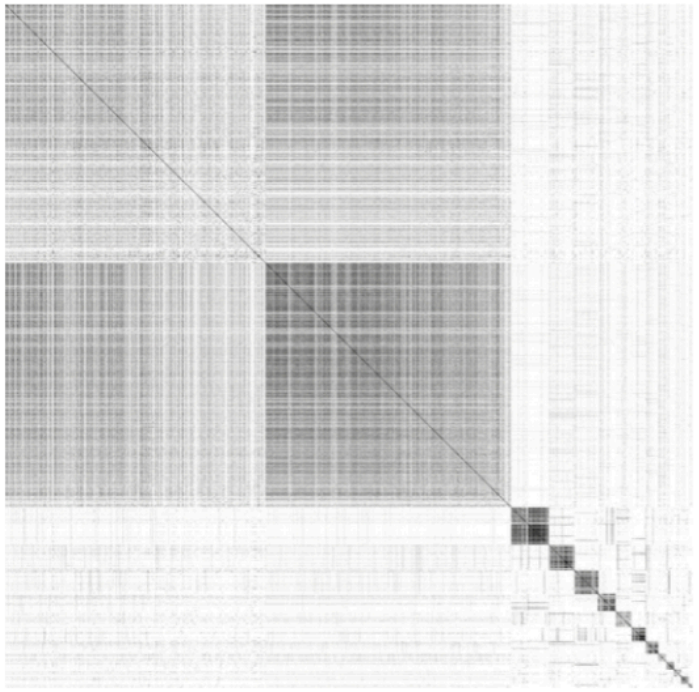


Fig. 13. The execution time of four clustering algorithms, e.g., PurTreeClust, HAC, SPEC and DBSCAN on 10 data sets listed in Table I.



(b) Clustering results by the PurTreeClust algorithm with $\gamma = 0.2$ and $k = 14$ on D_4.