多次元分類方式における木構造の構成の自動化

佐賀大学理工学部
知能情報システム学科
井ノ口　励　・　掛下　哲郎

研究の背景

情報量の増大 ➡ 情報を適切に分類する必要あり

階層構造を用いた分類
- 分類に一貫性なし

キーワードを用いた分類
- 情報の概観が不可能

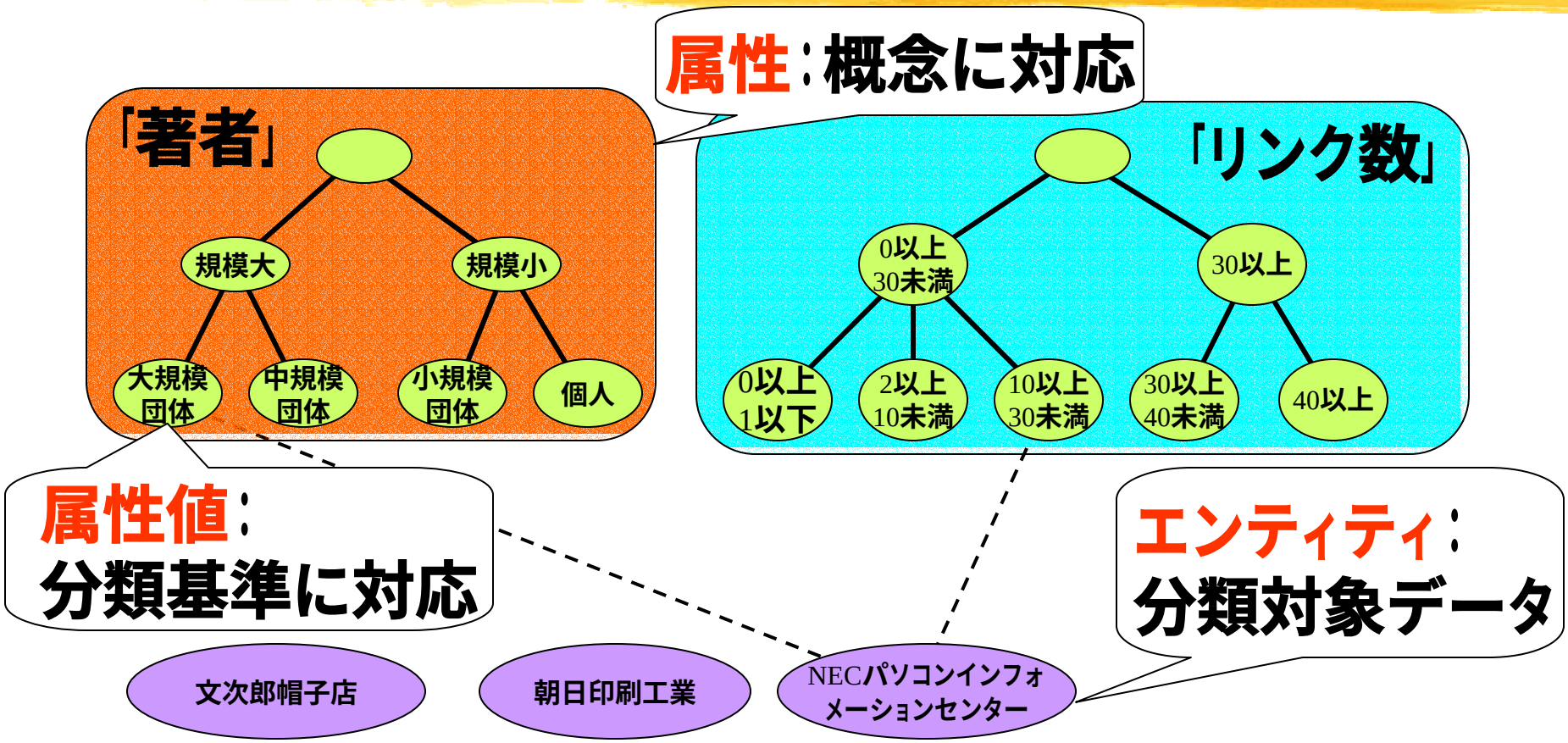
} 様々な欠点

⬇

◆ 多次元分類を用いた情報整理方式の提案
- ◆分類に一貫性あり
- ◆情報の概観が容易
- ◆検索の自由度が高い

本論文：木構造構成の自動化

多次元分類方式



木構造に要件 …… 様々な利点

多次元分類方式の要件と特徴

1. 属性値間に is-a / 排他関連 → 木構造に一貫性
2. 各木構造が平衡木
3. 中間属性値の子の数は $C_i/2$ 以上 C_i 以下

（2, 3 → 葉属性値指定の労力が均等化）

4. 葉属性値の対応エンティティ数は $m_i/2$ 以上 m_i 以下

（4 → エンティティ検索の労力が均等化）

様々な検索要求に公平に対応

木構造の構成の自動化

これまでの研究

- 多次元分類方式の提案
- 理論的分類と評価
- 実際のデータによる運用実験

木構造を手動で作成 → 管理者に大きな負担

本論文の目的：**木構造の構成の自動化**

1. 属性の候補の列挙
2. 適切な属性の抽出
3. 分類木の生成

} 自動化

自動化の流れ

(1).適切な属性の選択

1. 属性値数が少 …… 木構造が生成不可

   → 属性が持つ属性値数の調査

2. 属性間に従属性 …… 木構造の機能の損失

   → 属性間の独立性の調査

(2).属性の分類木を生成

属性の種類

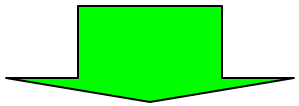
1. 属性値間に全順序がある場合
2. 属性値が位置情報の場合

} 木構造の自動生成アルゴリズム

3. 属性値が意味を持つ場合

適切な属性値数の調査

- 属性値：木構造のノードに対応
  属性値数が不十分　……木構造が極端に小規模



属性値数が少ない属性を候補から除外

(評価のデータ)会社四季報から取得した 406 社の企業

業種	資本金	所在地	設立年		従業員数	総合評価
44	303	87	85	37	387	5

属性値数が不十分

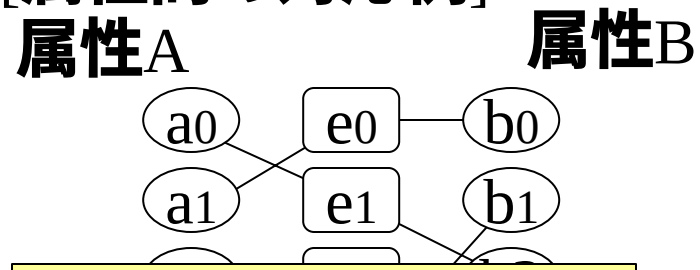
「総合評価」は属性として不適

属性間の独立性

属性間に従属性：検索機能の喪失 → 相関係数による独立性の評価

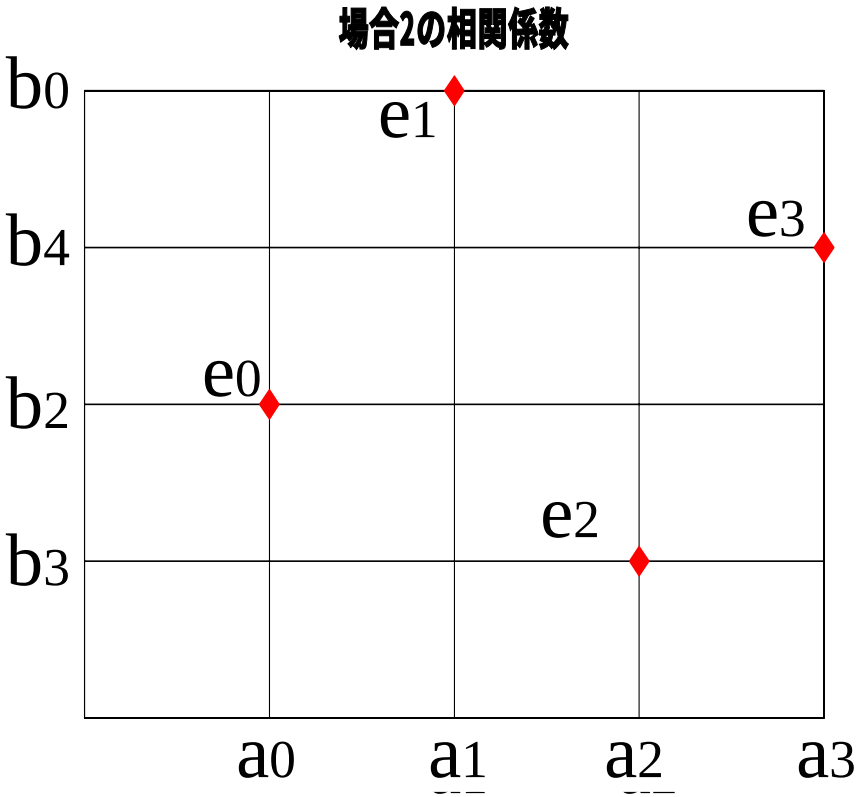
属性値と自然数の対応付けによって相関係数が変化

[属性間の対応例]



相関係数が高い：
属性間の独立性

	相関係数
場合1	1
場合2	0



属性間の独立性の評価

	資本金	所在地	設立年	主銀行
所在地	0.064901			
設立年	0.12205	0.05083		
主銀行	0.038		691	
従業員数	0.787371	0.029269	0.15067	0.04711

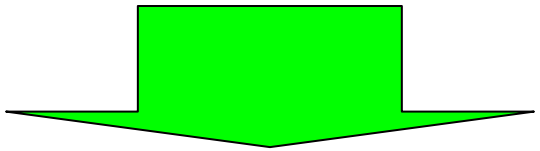
相関係数が大

「資本金」 ・・・・・ 企業の規模をより正確に表現

→ 属性として採用

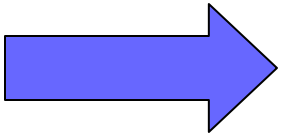
分類木の自動生成

分類木を手動で生成　……　労力大



分類木の自動生成

(1). 属性値間に全順序がある場合
(2). 属性値が位置情報である場合



分類木を自動生成する
アルゴリズムを2種類作成

属性値間に全順序がある場合

「資本金」や「設立年」が相当 …… 属性値間に大小関係

[アルゴリズム]

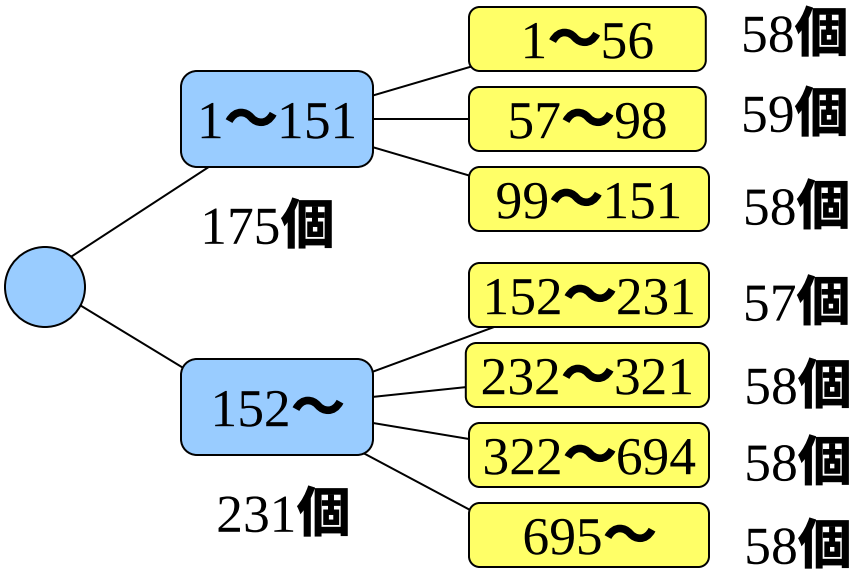
属性値

アルゴリズムによる分類木の生成

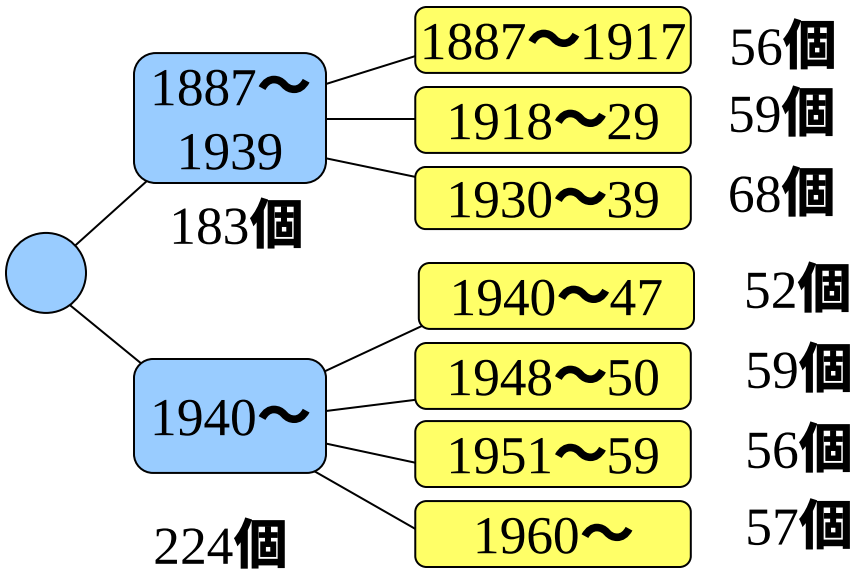
「資本金」と「設立年」の分類木を生成

(設定) 7 個のグループ , 各およそ 58 個の対応エンティティ

「資本金」の分類木



「設立年」の分類木



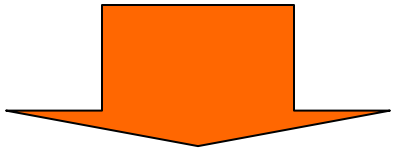
属性値が位置情報の場合

「所在地」が相当

→ グループに属性値を登録する方法が有効

属性値がスカラーで表現不可

→ 全順序がある場合のアルゴリズムは、適用不可



属性値を隣接するグループに登録

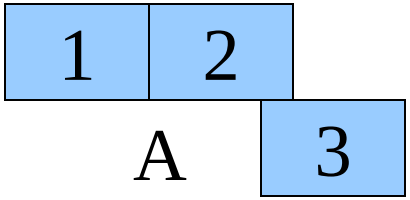
(属性値間の関連が強化)

位置情報を分類するアルゴリズム

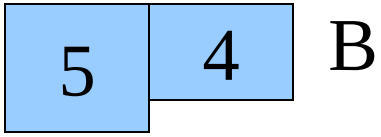
属性値間に大小関係なし ・・・・ 登録の順序が意味に非依存

2 種類のアルゴリズムを用意

[アルゴリズム 1]



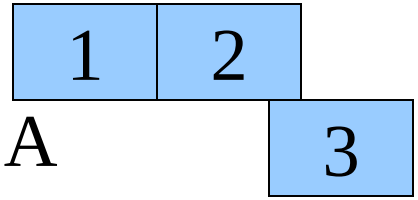
属性値



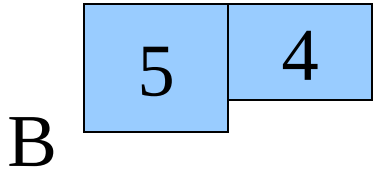
欠点：無所属の属性値が存在

補助処理：隣接グループに登録

[アルゴリズム 2]



属性値



欠点：要件を満たさない
グループが存在

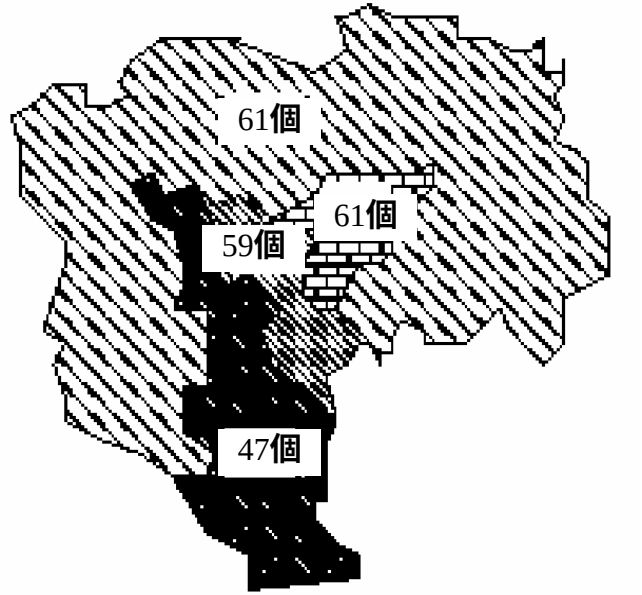
補助処理：グループ同士を合併

アルゴリズムによる木構造の生成

「所在地」の分類木を生成：東京 23 区の企業 228 社

(設定) 4 個のグループ , 約 58 個の対応エンティティ

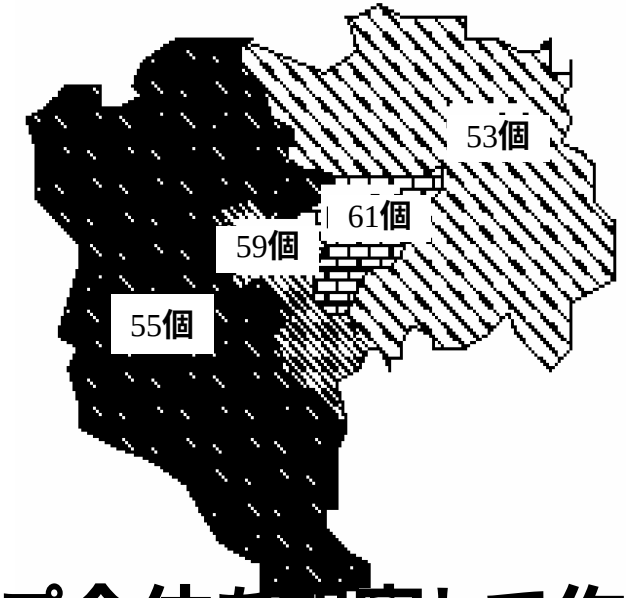
[アルゴリズム 1]



補助処理が属性値

→ 補助処理が容易

[アルゴリズム 2]



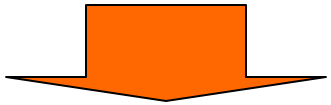
グループ全体を観察して作業

→ バランスが良好

まとめと今後の課題

多次元分類方式における木構造の構成の自動化

- 適切な属性値数を持つ属性の選択
- 属性間に独立性がある組み合わせの発見
- 全順序がある場合の分類木自動生成アルゴリズム
- 位置情報の場合の分類木自動生成アルゴリズム



属性値が数値で表現 → **自動生成可能**

今後の課題

- エンティティの意味を考慮した木構造構成の自動化

→ シソーラスの利用