最小サポート上昇法に基づく上位 k 関連パターン発見

亀谷由隆，佐藤泰介
東京工業大学

概要

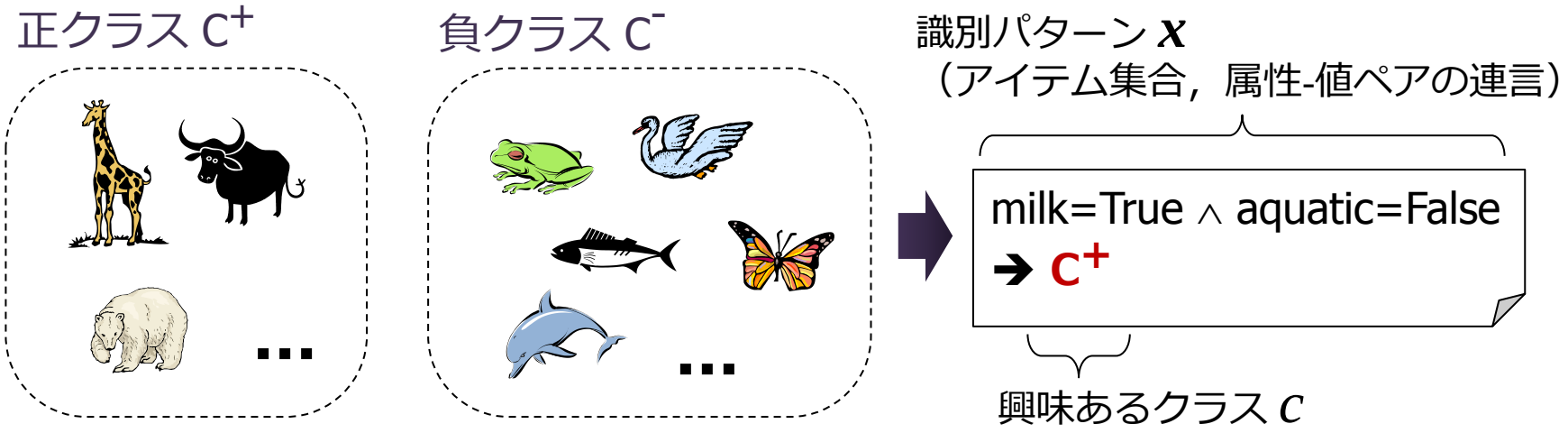
- 研究背景
- 提案手法 RP-growth
- 実験結果

概要

- 研究背景
- 提案手法 RP-growth
- 実験結果

背景：識別パターン発見 (1)

- 頻出パターン発見
  - 適切な最小サポートの設定が困難
  - 有益でないパターンが大量に出力される
- 改善法：
  - 上位 k パターン発見
  - 識別(discriminative)パターン発見
    - 識別パターン：クラス間の違いを際立たせる
    - 例：



背景：識別パターン発見 (2)

- 識別パターン発見
  - クラス情報を利用し，より意味のあるパターンを見つける
  - 幾つか異なる名前で呼ばれている：
    - Emerging pattern mining [Dong & Li 99]
    - Contrast set mining [Bay & Pazzani 01]
    - Subgroup discovery [Wrobel 97]
    - Supervised descriptive rule discovery [Kralj Novak et al. 09]
    - Cluster grouping [Zimmermann & De Raedt 09]
    - …
  - 応用
    - 高精度・理解可能性を備えた分類器の構築
    - 変化検出
    - クラスタリング結果のラベリング（自動的特徴づけ）
    - ….

背景: 関連スコア

規則 $\boldsymbol{x}$ ➔ c における興味あるクラス c とパターン $\boldsymbol{x}$ の関連 (relevance) を測る尺度

- 正例でのサポート（再現率）$p(\boldsymbol{x} \mid c)$
- 確信度（精度）$p(c \mid \boldsymbol{x}) \propto \frac{p(\boldsymbol{x} \mid c)}{p(\boldsymbol{x})}$
- F値 $\mathrm{F}_c(\boldsymbol{x}) = \frac{2p(c \mid \boldsymbol{x})p(\boldsymbol{x} \mid c)}{p(c \mid \boldsymbol{x}) + p(\boldsymbol{x} \mid c)} \propto \frac{p(\boldsymbol{x} \mid c)}{p(c) + p(\boldsymbol{x})}$
- χ^2値 $\chi^2_c(\boldsymbol{x}) = \sum_{c' \in \{c, \neg c\},\, \boldsymbol{x}' \in \{\boldsymbol{x}, \neg \boldsymbol{x}\}} \tau(c', \boldsymbol{x}')$

  $\tau(c', \boldsymbol{x}') = N \frac{p(c', \boldsymbol{x}') - p(c')p(\boldsymbol{x}')}{p(c')p(\boldsymbol{x}')}$
- Support difference [Bay & Pazzani 01] $\mathrm{SupDiff}_c(\boldsymbol{x}) \stackrel{\mathrm{def}}{=} p(\boldsymbol{x} \mid c) - p(\boldsymbol{x} \mid \neg c)$

背景: 分岐限定法

- これまで挙げてきた関連スコアの多くはパターン間の包含関係に関する反単調性 (anti-monotonicity) を満たさない
  - Apriori 等で行っているような枝刈りをそのまま適用できない
- 分岐限定法 (branch-and-bound):
  - 関連スコアの上界を計算する
  - その上界に基づき枝刈りを行う
- 既存研究:
  - Subgroup discovery [Wrobel 97]
  - AprioriSMP [Morishita & Sese 00]
  - CG アルゴリズム [Zimmermann & De Raedt 09]

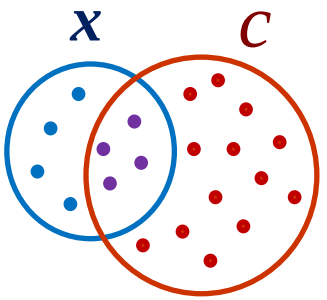
関連スコアの凹性 (convexity) を利用
(情報利得, categoriy utility, χ^2)

概要

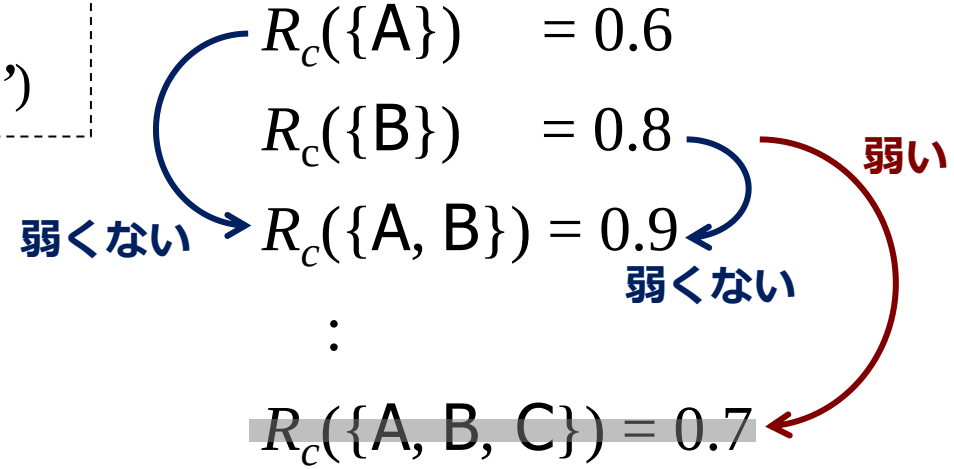
- ✓研究背景
- 提案手法 RP-growth
- 実験結果

提案手法: RP-growth

- 興味あるクラス c に対し，関連スコア R_c の上位 k パターンを見つける
- 制約（$\boldsymbol{x}$: パターン）
  - サポート $p(\boldsymbol{x} \mid c) \geq \sigma_{\min}$ （デフォルト $\sigma_{\min} = 1/|D|$）
  - 確信度 $p(c \mid \boldsymbol{x}) \geq \beta_{\min}$ （デフォルト $\beta_{\min} = 0.5$ or $p(c)$）
  - 2つのパターン $\boldsymbol{x}$ と $\boldsymbol{x}'$ は互いに弱くない

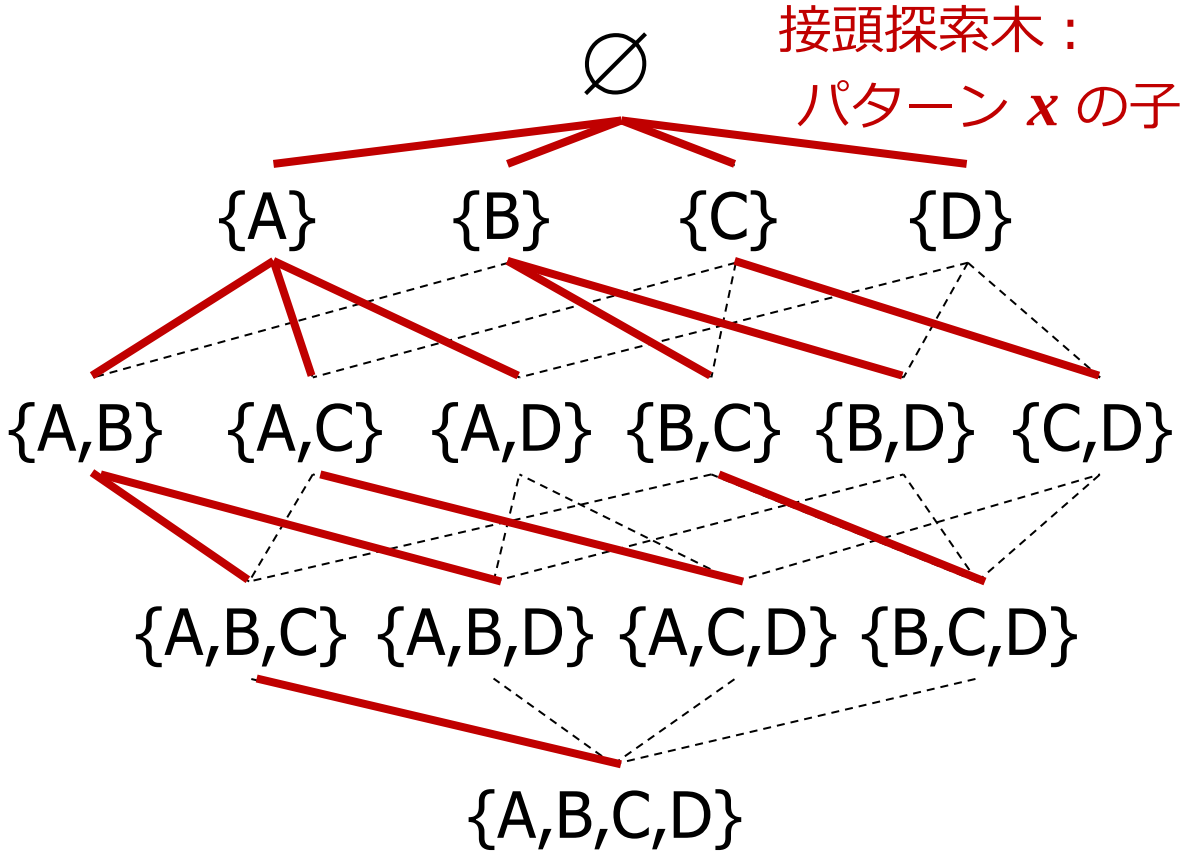


「$\boldsymbol{x}'$ は $\boldsymbol{x}$ より弱い」$\Leftrightarrow$
$\boldsymbol{x} \subset \boldsymbol{x}'$ にも関わらず $R_c(\boldsymbol{x}) \geq R_c(\boldsymbol{x}')$



頻出パターン発見 (1)

- パターン空間上の束



接頭探索木：

パターン $\boldsymbol{x}$ の子 = $\boldsymbol{x} \cup \{i\}$

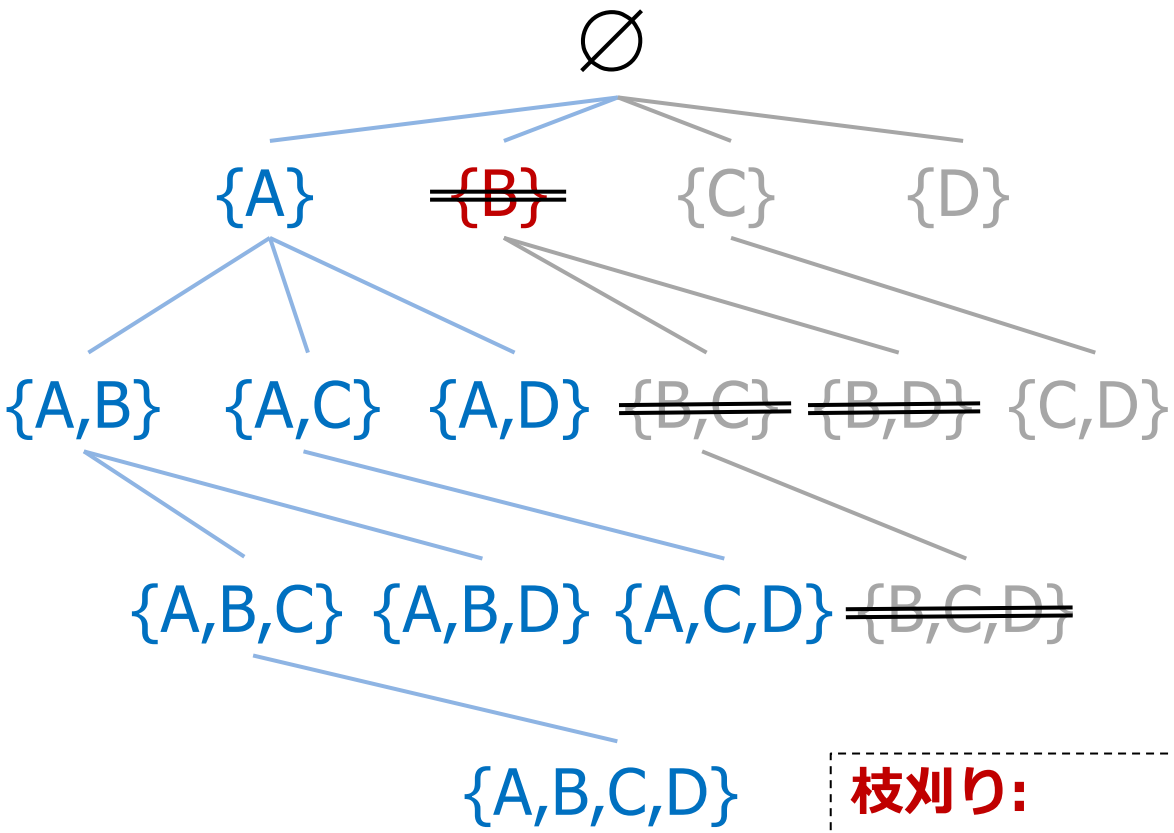
s.t. $i > j,\ j \in \boldsymbol{x}$

A, B, C, D: 属性-値の対（アイテム）

全順序 A < B < C < D

頻出パターン発見 (2)

- 最小サポート σ_{min}を指定: 深さ優先探索 + Apriori 流の枝刈り

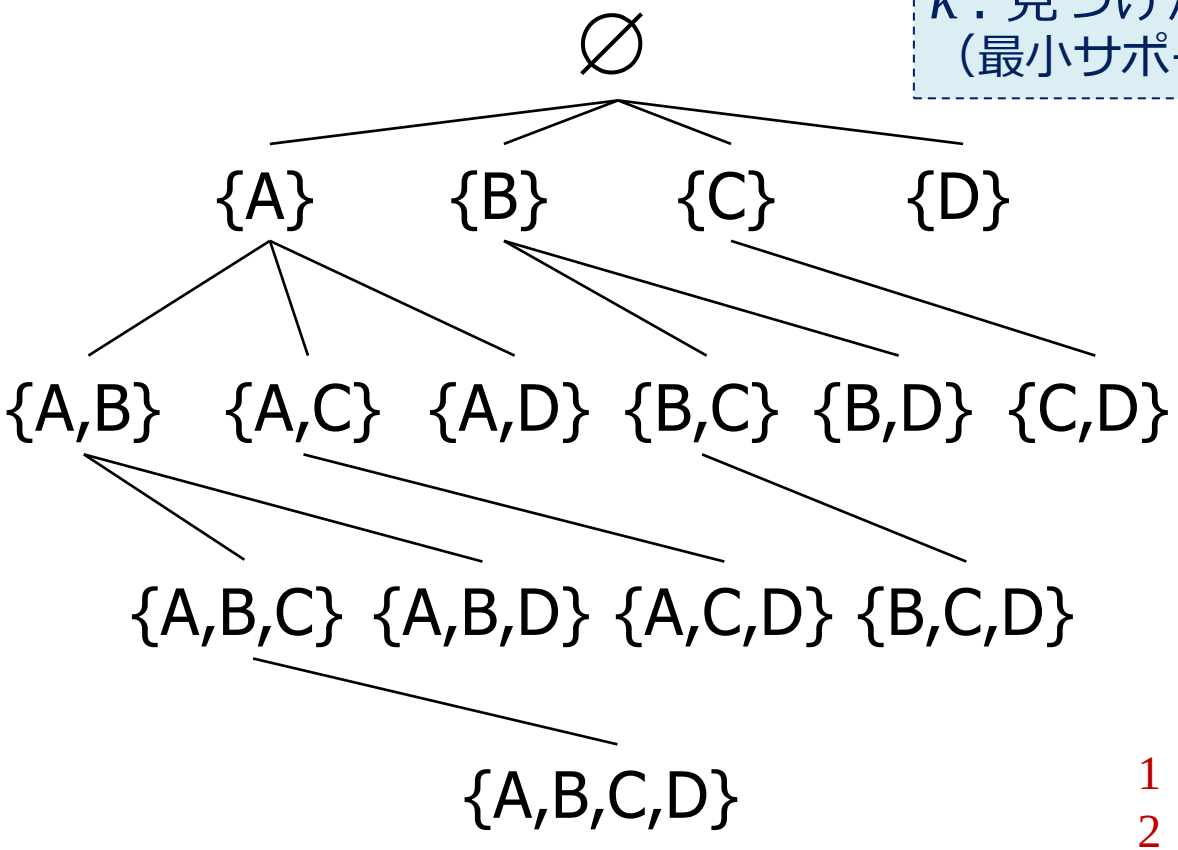


枝刈り:
$p(\{B\} \mid c) < \sigma_{min}$ なら
{B,C}, {B,C,D}, ... を訪問しない

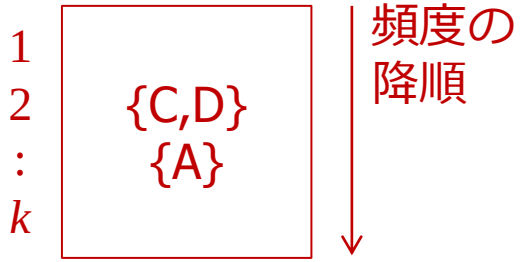
上位 k 頻出パターン発見 (1)

- **方針**: 深さ優先探索 + 最小サポート上昇 (minimum support raising)

k : 見つけたいパターン数
（最小サポートより指定が容易）



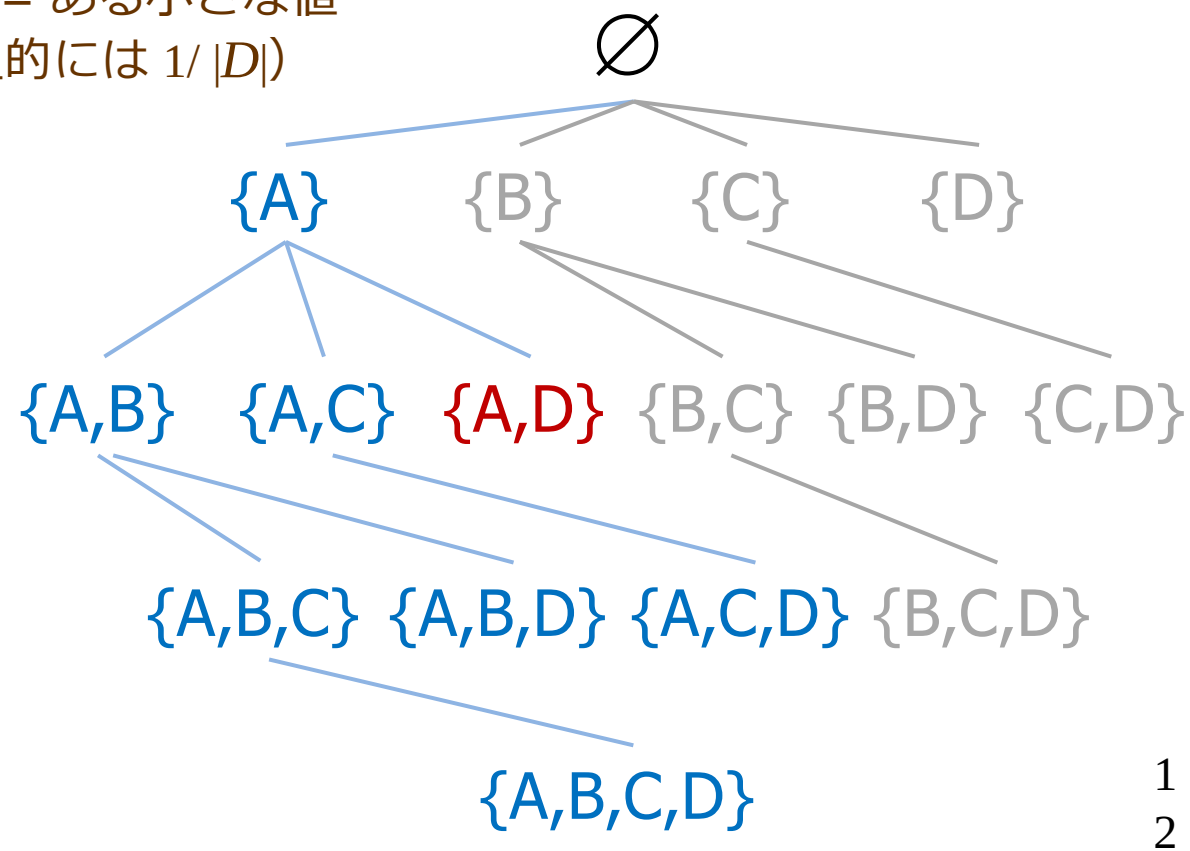
候補リスト



上位 k 頻出パターン発見 (2)

- **方針**: 深さ優先探索 + 最小サポート上昇 (minimum support raising)

σ_{min} := ある小さな値
(典型的には $1/|D|$)



候補リスト

1	
2	
:	
k	{A,D}

最小サポート上昇: $\sigma_{min} := p(\{A, D\} \mid c)$

RP-growth:上位 k 関連パターン発見

- **方針**: 分岐限定法
  - F 値の定義:

$$\mathrm{F}_c(\boldsymbol{x}) = \frac{2p(\boldsymbol{x} \mid c)p(c \mid \boldsymbol{x})}{p(\boldsymbol{x} \mid c) + p(c \mid \boldsymbol{x})}$$

  - $p(c \mid \boldsymbol{x}) := 1$ と代入し，反単調性を満たす $\mathrm{F}_c(\boldsymbol{x})$ の上界を得る

$$\overline{\mathrm{F}}_c(\boldsymbol{x}) = \frac{2p(\boldsymbol{x} \mid c)}{p(\boldsymbol{x} \mid c) + 1}$$

  - 枝刈り: 以下の場合，$\boldsymbol{x}$ を含むパターンは候補リストに残らない：

$$\mathrm{F}_c(\boldsymbol{z}) > \overline{\mathrm{F}}_c(\boldsymbol{x}) = \frac{2p(\boldsymbol{x} \mid c)}{p(\boldsymbol{x} \mid c) + 1} \quad \overset{\text{iff}}{\longleftrightarrow} \quad p(\boldsymbol{x} \mid c) < \frac{\mathrm{F}_c(\boldsymbol{z})}{2 - \mathrm{F}_c(\boldsymbol{z})}$$

    $\boldsymbol{z}$: k 番目のパターン

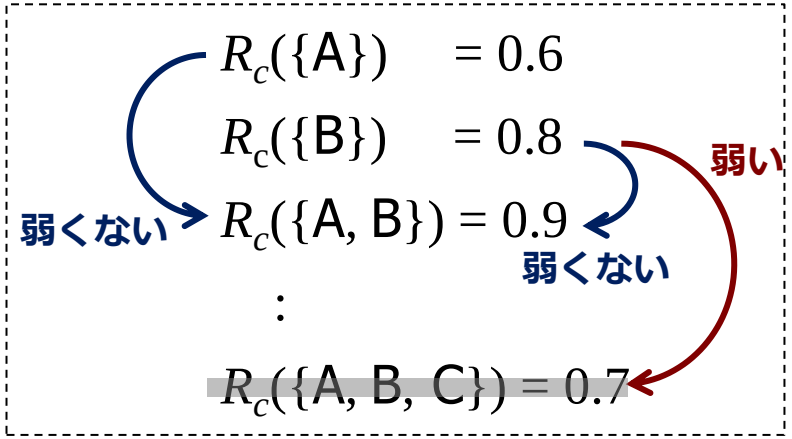
  - 最小サポート上昇:

$$\sigma_{\min} := \frac{\mathrm{F}_c(\boldsymbol{z})}{2 - \mathrm{F}_c(\boldsymbol{z})}$$

- 凹性を満たさない関連スコアにも適用可能
  （F値，Jaccard 係数，support difference）
- データベース縮約などの高速化技法が使える
- どんな関連スコアでもうまくいく訳ではない
  （確信度, growth rate, 情報利得, TF-IDF）

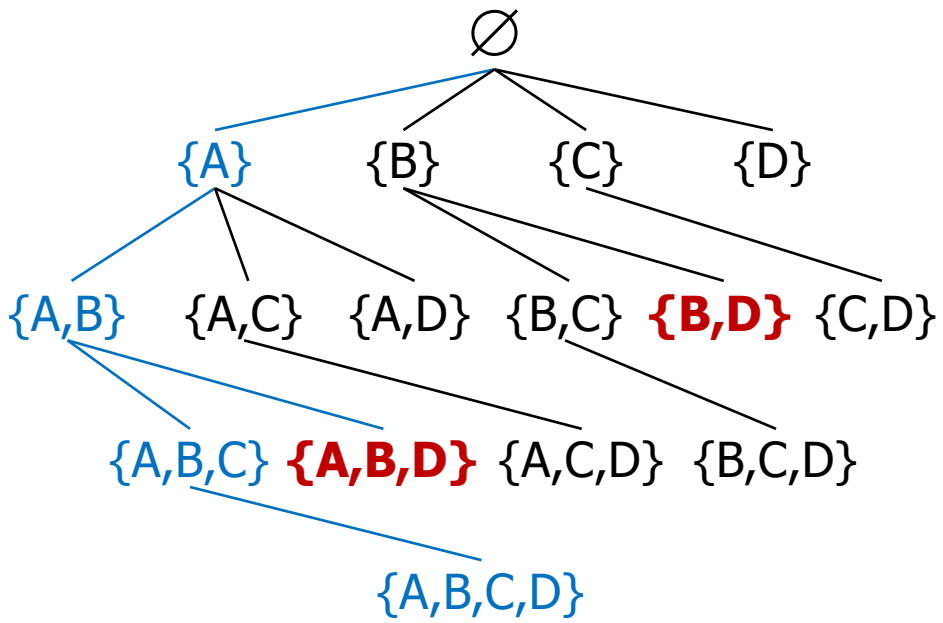
「弱い」パターンの扱い

- RP-growth:
  - 「弱い」パターンの検査
  - 分岐限定法の考えに基づく「弱い」パターンの枝刈り



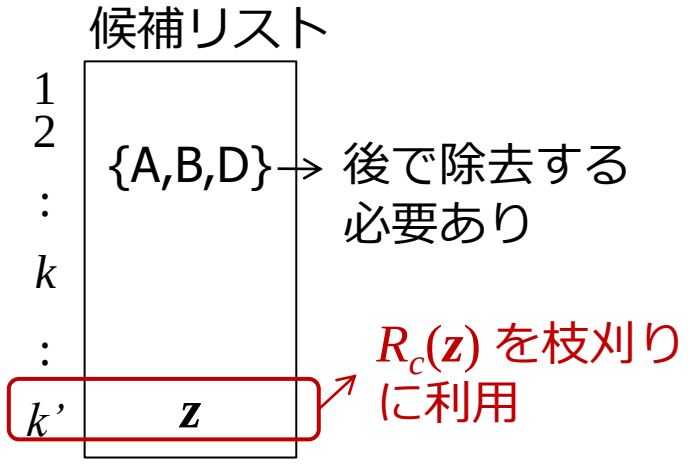
「弱い」パターンの検査 (1)

- 「弱い」の検査ではパターン間の包含関係の検査が必要
- 幅優先探索: 包含検査は容易だが一般にメモリ消費量大
- 深さ優先探索: 接頭探索木では単純ではない
  - パターン x の部分パターンが x の後に見つかる
  - 「弱い」パターンを暫定的に候補リストに入れると最小サポート上昇の効果が落ちる



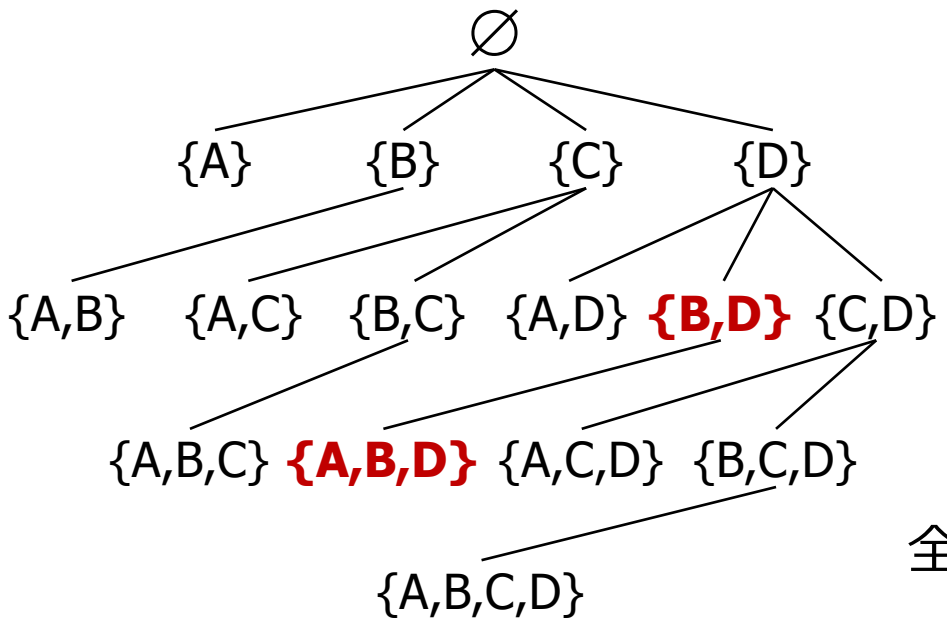
例：

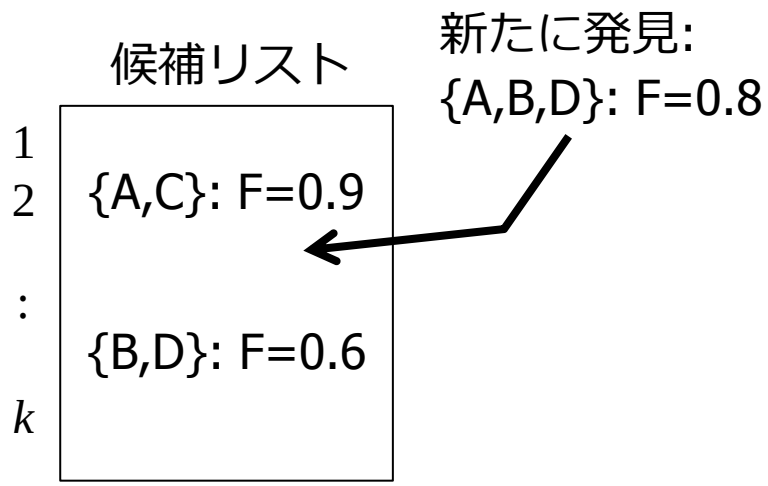
$R_c(\{A,B,D\}) \leq R_c(\{B,D\})$



「弱い」パターンの検査 (2)

- 接尾探索木の導入
  - パターン $\boldsymbol{x}$ の子 = $\boldsymbol{x} \cup \{i\}$ s.t. $i < j, j \in \boldsymbol{x}$
  - 「$\boldsymbol{x}$ を訪問する時点で $\boldsymbol{x}$ の部分パターンは訪問済み」
  - パターン追加時に候補リストの中だけで検査すれば充分
  - FP-growth では（暗黙のうちに）接尾探索木を使用

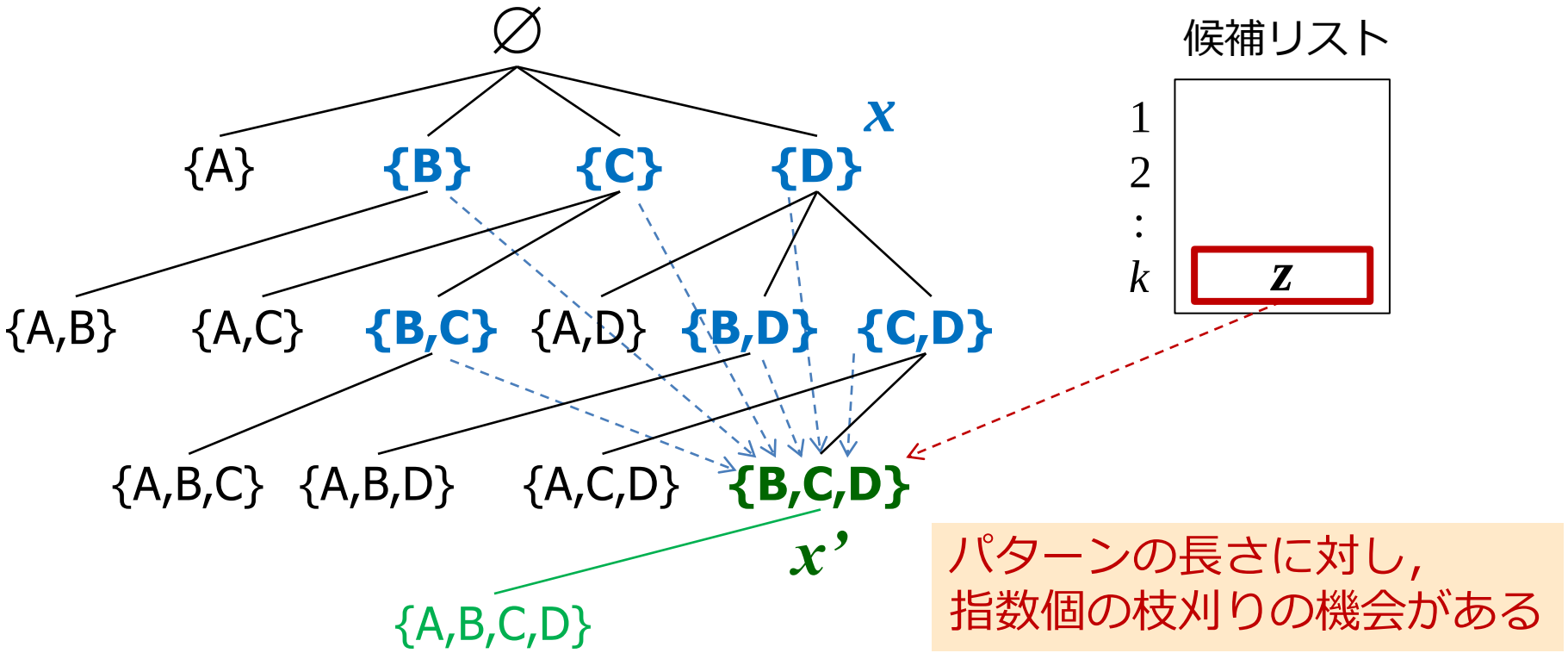




全順序 A < B < C < D

「弱い」パターンの枝刈り

- パターン x' とその部分パターン x
- 上界 $\overline{R}_c(x') \leq R_c(x)$ であるとき，x' を含むパターンは全て x より弱いことが保証 ➔ x' 以下は枝刈り可能



概要

- ✓研究背景
- ✓提案手法 RP-growth
- 実験結果

実験

- 20 newsgroup データ
  - J. Rennie による前処理済みデータに更に前処理を行う
  - 5,666 単語から成る 17,930 記事
- タスク
  - 上位25関連パターンの発見
  - テキスト分類（関連パターンに基づく属性構築）

実験：得られた上位 25 関連パターン (1)

- ニュースグループ comp.graphics
- 関連スコア：F 値
- 制約 $p(c \mid \boldsymbol{x}) \geq 0.5$
- 画像処理関連の単語（の組）が抽出
- 確信度（精度）$p(c \mid \boldsymbol{x})$ とサポート（再現率）$p(\boldsymbol{x} \mid c)$ から各パターンの出現の様子を観察できる

Pattern $\boldsymbol{x}$	$p(c \mid \boldsymbol{x})$	$p(\boldsymbol{x} \mid c)$	$\mathrm{F}_c(\boldsymbol{x})$
{graphic, program}	0.537	0.136	0.217
{gif}	0.552	0.119	0.196
{graphic, imag}	0.642	0.108	0.185
{imag, program}	0.516	0.110	0.181
{imag, file}	0.531	0.105	0.175
{graphic, find}	0.578	0.087	0.151
{imag, bit}	0.514	0.083	0.144
{graphic, code}	0.613	0.081	0.143
{graphic, bit}	0.545	0.080	0.140
{graphic, packag}	0.591	0.076	0.134
{format, convert}	0.588	0.075	0.132
{graphic, comp}	0.730	0.072	0.132
{imag, format}	0.613	0.072	0.129
{graphic, point}	0.573	0.070	0.125
{graphic, format}	0.670	0.068	0.123
{imag, convert}	0.596	0.066	0.118
{polygon}	0.915	0.060	0.113
{imag, softwar}	0.500	0.062	0.111
{graphic, ftp}	0.500	0.061	0.109
{graphic, algorithm}	0.852	0.058	0.108
{jpeg}	0.825	0.058	0.108
{graphic, group}	0.514	0.060	0.108
{graphic, site}	0.530	0.059	0.106
{graphic, comput, articl}	0.525	0.059	0.106
{code, algorithm}	0.500	0.059	0.105

実験：得られた上位 25 関連パターン (2)

- ニュースグループ rec.sport.hockey
- ホッケー関連の単語（の組）が抽出

Pattern $\boldsymbol{x}$	$p(c \mid \boldsymbol{x})$	$p(\boldsymbol{x} \mid c)$	$\mathrm{F}_c(\boldsymbol{x})$
{hockei}	0.943	0.377	0.538
{team}	0.519	0.473	0.495
{playoff}	0.943	0.277	0.428
{game, plai}	0.506	0.273	0.354
{nhl}	0.990	0.206	0.341
{cup}	0.584	0.195	0.292
{player, plai}	0.575	0.190	0.286
{score}	0.510	0.194	0.281
{game, player}	0.561	0.186	0.280
{game, goal}	0.899	0.157	0.267
{game, win}	0.517	0.174	0.260
{game, fan}	0.622	0.164	0.260
{plai, goal}	0.852	0.144	0.246
{wing}	0.515	0.156	0.240
{leaf}	0.894	0.132	0.230
{bruin}	1.000	0.130	0.230
{pittsburgh}	0.567	0.142	0.226
{game, watch}	0.621	0.136	0.224
{detroit}	0.733	0.131	0.222
{penguin}	0.871	0.127	0.222
{game, season}	0.539	0.137	0.219
{game, night}	0.660	0.129	0.216
{ranger}	0.629	0.129	0.214
{plai, win}	0.529	0.134	0.214
{plai, fan}	0.603	0.128	0.211

実験：得られた上位 25 関連パターン (3)

- ニュースグループ
  talk.politics.guns
- 銃関連の単語（の組）が抽出

Pattern $\boldsymbol{x}$	$p(c \mid \boldsymbol{x})$	$p(\boldsymbol{x} \mid c)$	$\mathrm{F}_c(\boldsymbol{x})$
{gun}	0.540	0.414	0.469
{weapon}	0.528	0.253	0.342
{fbi}	0.506	0.246	0.331
{firearm}	0.884	0.196	0.321
{batf}	0.662	0.155	0.252
{waco}	0.543	0.154	0.240
{assault}	0.587	0.124	0.205
{cdt, sw}	0.933	0.110	0.196
{cdt, stratu}	0.916	0.110	0.196
{handgun}	0.818	0.111	0.195
{cdt}	0.817	0.110	0.193
{stratu, sw}	0.700	0.110	0.190
{fire, compound}	0.698	0.109	0.188
{stratu}	0.570	0.110	0.184
{bd}	0.530	0.110	0.182
{sw}	0.521	0.110	0.181
{atf}	0.692	0.101	0.176
{arm, law}	0.527	0.086	0.148
{compound, dai}	0.598	0.082	0.144
{nra}	0.696	0.079	0.143
{rocket, special}	0.750	0.077	0.140
{rocket, speak}	0.840	0.076	0.139
{rocket, vo}	0.918	0.075	0.139
{vo, investor}	0.918	0.075	0.139
{vo, speak, todai}	0.918	0.075	0.139

実験：テキスト分類

**: 最良
*: ２番目に良い

- 分類器 SVM（実装：LIBSVM）
- 関連パターンに基づく合成属性を導入すると線形カーネルでも RBF カーネルに近い分類性能が得られる（関連スコアによる差は見られず）

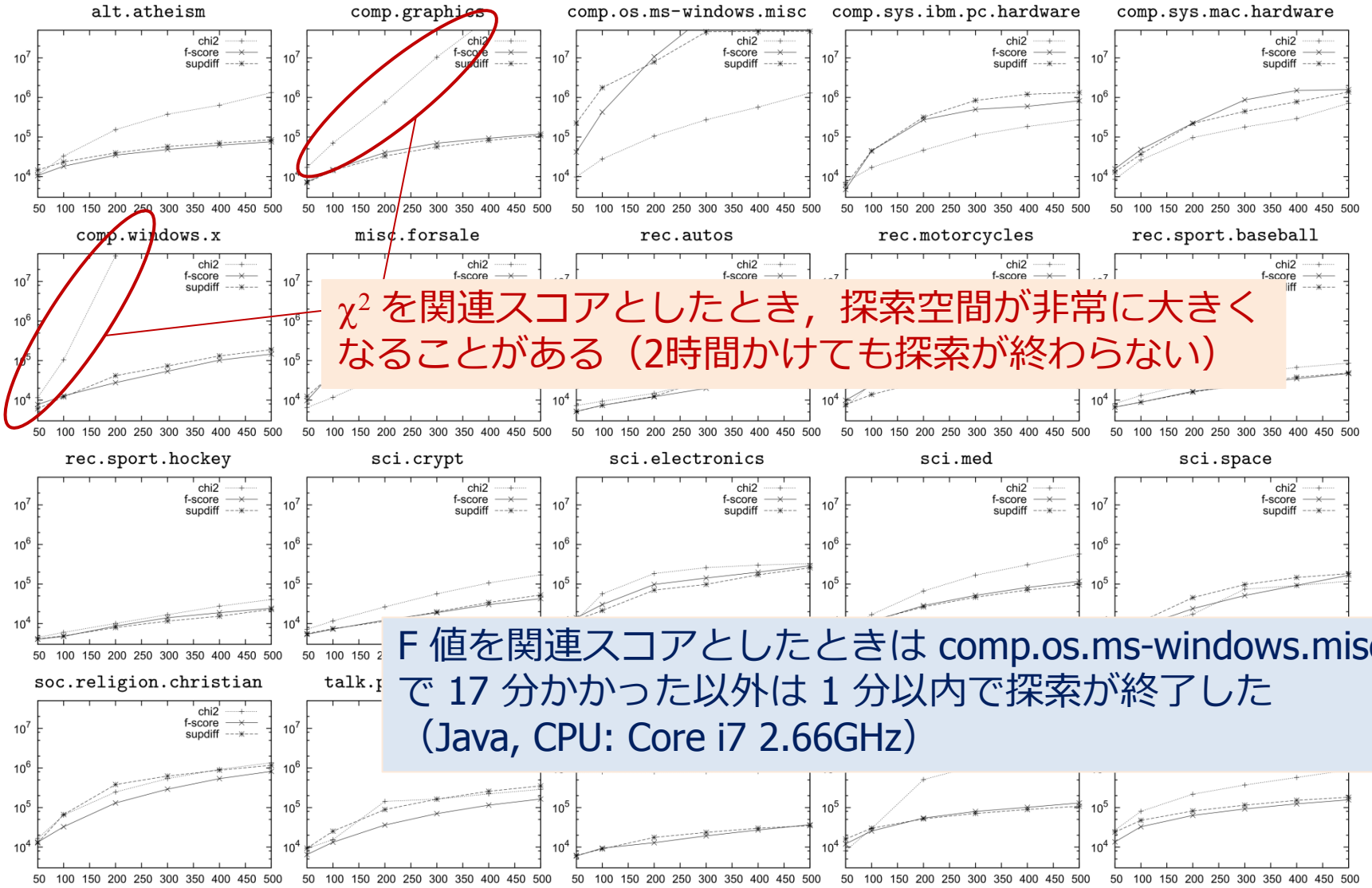
ニュースグループ	単体属性のみ		単体属性と合成属性の混合		
	線形カーネル	RBF カーネル	線形カーネル		
			χ^2	F 値	Support diff.
alt.atheism	**83.97±1.39	*83.78±1.37	83.68±1.14	83.68±1.42	83.76±1.41
comp.graphics	71.18±0.97	**72.96±1.13	N/A	*72.02±0.79	71.98±0.87
comp.os.ms-windows.misc	75.73±1.28	**77.57±1.03	75.84±1.20	76.17±1.31	*76.25±1.28
comp.sys.ibm.pc.hardware	68.75±1.89	*68.93±1.34	68.45±1.32	**69.26±1.60	68.92±1.80
comp.sys.mac.hardware	*79.69±1.23	**80.85±1.29	79.63±1.60	78.72±1.37	78.68±1.43
comp.windows.x	80.65±1.04	*80.68±1.20	N/A	80.63±1.15	**81.06±1.23
misc.forsale	79.59±1.17	**80.08±0.96	N/A	*79.85±1.23	79.83±1.12
rec.autos	81.20±1.16	**82.29±1.38	81.15±1.29	81.24±1.12	*81.47±1.24
rec.motorcycles	91.22±0.53	91.06±0.48	*91.70±0.46	91.67±0.57	**91.70±0.53
rec.sport.baseball	90.14±0.51	90.49±0.54	**90.72±0.52	90.39±0.48	*90.51±0.45
rec.sport.hockey	94.35±0.50	94.44±0.50	94.68±0.47	*94.89±0.53	**94.95±0.54
sci.crypt	91.27±0.60	91.31±0.64	**91.63±0.58	*91.61±0.56	91.39±0.61
sci.electronics	71.27±0.89	**73.65±1.58	71.56±1.00	73.21±0.99	*73.35±0.86
sci.med	85.98±0.78	*86.59±0.70	86.42±0.72	**86.60±0.70	86.42±0.74
sci.space	89.75±0.55	90.22±0.57	**90.85±0.43	*90.54±0.44	90.49±0.47
soc.religion.christian	85.41±0.65	**86.12±0.64	*86.12±0.66	86.03±0.55	85.90±0.58
talk.politics.guns	83.19±0.93	**84.62±1.10	*84.21±1.01	83.89±1.15	83.82±1.17
talk.politics.mideast	*92.65±0.72	**92.70±0.72	92.33±0.65	92.22±0.66	92.10±0.67
talk.politics.misc	76.34±1.23	**77.28±1.13	76.10±1.13	*76.96±1.25	76.62±1.50
talk.religion.misc	62.61±1.44	*62.75±1.50	**63.50±1.84	61.89±2.00	62.52±1.71
All	83.88±0.20	**84.95±0.22	84.48±0.13	84.73±0.22	*84.73±0.23

2クラス分類

多クラス分類

実験：探索効率

- 上位 k パターンの発見までに訪問したパターン数（x 軸: k）



χ^2 を関連スコアとしたとき，探索空間が非常に大きくなることがある（2時間かけても探索が終わらない）

F 値を関連スコアとしたときは comp.os.ms-windows.misc で 17 分かかった以外は 1 分以内で探索が終了した（Java, CPU: Core i7 2.66GHz）

まとめ

- 最小サポート上昇に基づく上位 k 関連パターン発見手法 RP-growth
  - 凹性を満たさない関連スコアに対する分岐限定法
  - 最小サポート上昇への翻訳（データベース縮約）
  - 「弱い」関連パターンと枝刈り（接尾探索木の導入）

今後の課題

- スケーラビリティの調査
- より tight な上界の探究
- 系列，木，グラフへの適用
- パターンの表現力の向上（選言，概念階層，述語論理の導入）
- 応用：
  - ベイジアンネットの説明的分析
  - 属性再構築（= 属性選択 + 属性構築）
  - …

背景: 関連スコア (1)

規則 x ➔ c における興味あるクラス c とパターン x の関連 (relevance) を測る尺度

- 正例でのサポート（再現率）
- 確信度（精度）
- Growth rate [Dong & Li 99]
- Lift [Geng & Hamilton 06]
- 点相互情報量 [Church & Hanks 89]

$$p(\boldsymbol{x} \mid c)$$

パターンに同じ順序を与える [Kralj Novak et al. 09]

$$p(c \mid \boldsymbol{x}) \propto \frac{p(\boldsymbol{x} \mid c)}{p(\boldsymbol{x})}$$

$$\mathrm{GR}_c(\boldsymbol{x}) \stackrel{\mathrm{def}}{=} \frac{p(\boldsymbol{x} \mid c)}{p(\boldsymbol{x} \mid \neg c)}$$

$$\mathrm{Lift}_c(\boldsymbol{x}) \stackrel{\mathrm{def}}{=} \frac{p(c, \boldsymbol{x})}{p(c)p(\boldsymbol{x})} = \frac{p(\boldsymbol{x} \mid c)}{p(\boldsymbol{x})}$$

$$\mathrm{PMI}_c(\boldsymbol{x}) \stackrel{\mathrm{def}}{=} \log \frac{p(c, \boldsymbol{x})}{p(c)p(\boldsymbol{x})}$$

背景: 関連スコア (2)

規則 $\boldsymbol{x}$ ➔ c における興味あるクラス c とパターン $\boldsymbol{x}$ の関連 (relevance) を測る尺度

- Leverage [Geng & Hamilton 06]

$$\mathrm{Leverage}_c(\boldsymbol{x}) \stackrel{\mathrm{def}}{=} p(c, \boldsymbol{x}) - p(c)p(\boldsymbol{x})$$

- Support difference [Bay & Pazzani 01]

$$\mathrm{SupDiff}_c(\boldsymbol{x}) \stackrel{\mathrm{def}}{=} p(\boldsymbol{x} \mid c) - p(\boldsymbol{x} \mid \neg c)$$

- Weighted relative accuracy [Wrobel 97]

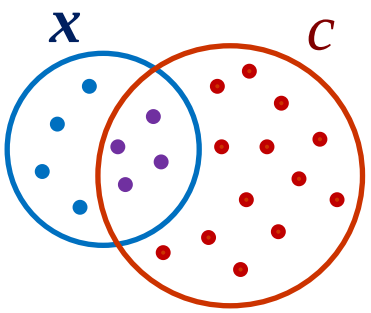
$$\begin{aligned} \mathrm{WRAcc}_c(\boldsymbol{x}) &\stackrel{\mathrm{def}}{=} p(\boldsymbol{x})(p(c \mid \boldsymbol{x}) - p(c)) \\ &\propto p(\boldsymbol{x} \mid c) - p(\boldsymbol{x}) \end{aligned}$$

パターンに同じ順序を与える
[Kralj Novak et al. 09]

背景: 関連スコア (3)

規則 x ➔ c における興味あるクラス c とパターン x の関連 (relevance) を測る尺度

- F値 $\mathrm{F}_c(\boldsymbol{x}) = \dfrac{2p(c \mid \boldsymbol{x})p(\boldsymbol{x} \mid c)}{p(c \mid \boldsymbol{x}) + p(\boldsymbol{x} \mid c)} \propto \dfrac{p(\boldsymbol{x} \mid c)}{p(c) + p(\boldsymbol{x})}$
- Jaccard 係数 $\mathrm{J}_c(\boldsymbol{x}) = \dfrac{p(c, \boldsymbol{x})}{p(c) + p(\boldsymbol{x}) - p(c, \boldsymbol{x})}$
- TF-IDF $\mathrm{TF\text{-}IDF}_c(\boldsymbol{x}) = p(\boldsymbol{x} \mid c) \log \dfrac{1}{p(\boldsymbol{x})}$



パターンに同じ順序を与える

cf. 確信度（精度）

$$p(c \mid \boldsymbol{x}) \propto \frac{p(\boldsymbol{x} \mid c)}{p(\boldsymbol{x})}$$

背景: 関連スコア (4)

規則 $\boldsymbol{x}$ ➔ c における興味あるクラス c とパターン $\boldsymbol{x}$ の関連 (relevance) を測る尺度

- χ^2値

$$\chi^2_c(\boldsymbol{x}) = \sum_{c' \in \{c, \neg c\},\ \boldsymbol{x}' \in \{\boldsymbol{x}, \neg \boldsymbol{x}\}} \tau(c', \boldsymbol{x}')$$

$$\tau(c', \boldsymbol{x}') = N \frac{p(c', \boldsymbol{x}') - p(c')p(\boldsymbol{x}')}{p(c')p(\boldsymbol{x}')}$$

- 情報利得 (information gain)
- Gini index
- Category utility（COBWEB で使用）

関連スコア $R_c(x)$ の上界

規則 $x \Rightarrow c$

- まず c と x の分割表を考える

	c	$\neg c$
x	$p(c, x)$	$p(\neg c, x)$
$\neg x$	$p(c, \neg x)$	$p(\neg c, \neg x)$

大きくすれば $R_c(x)$ は向上
小さくすれば $R_c(x)$ は向上

- x を拡大して x' としたときを考える

	c	$\neg c$
x	$p(c, x)$	$p(\neg c, x)$
$\neg x$	$p(c, \neg x)$	$p(\neg c, \neg x)$

x' に拡大すると元より小さくなる
x' に拡大すると元より大きくなる

※ 2番目の分割表で示した性質より，
x を x' に拡大し $p(c, \neg x') = 0$ とできない

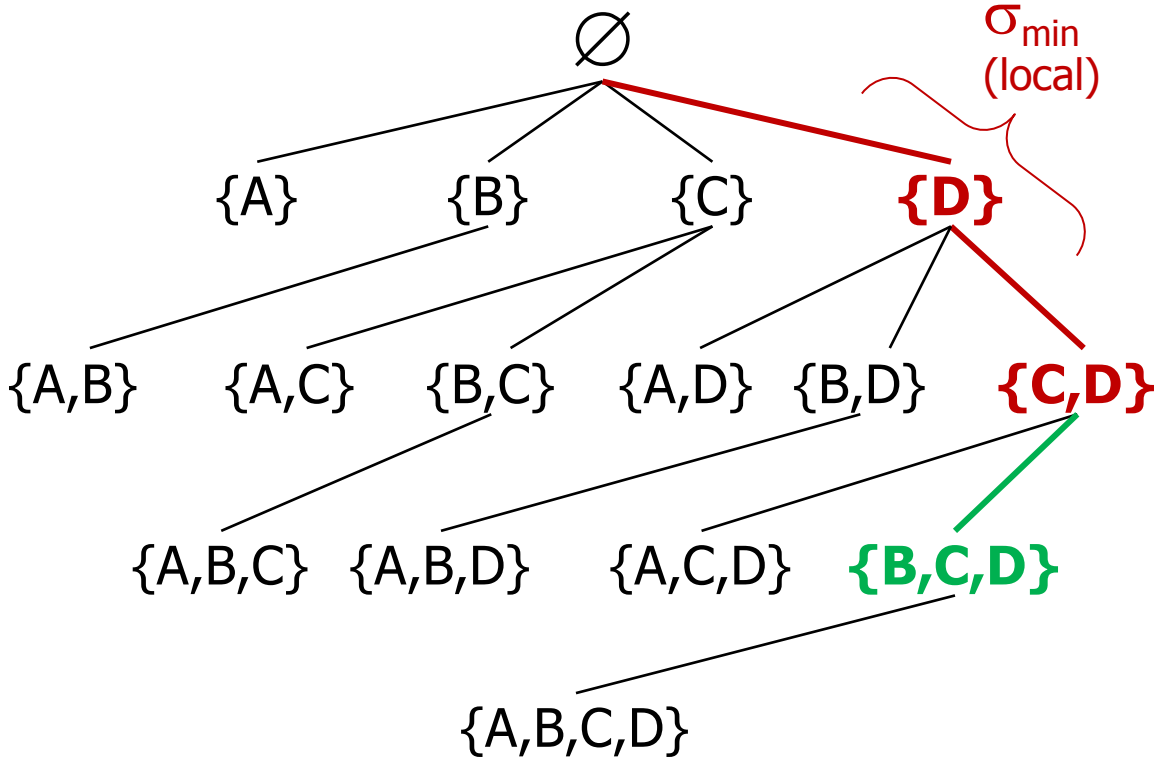
- **楽観的予測**：
「x を x' に拡大したとき $p(\neg c, x') = 0$ となる x' がある」
➔ $R_c(x)$ にて $p(\neg c, x) := 0$ とおいたときの値は $R_c(x)$ の上界

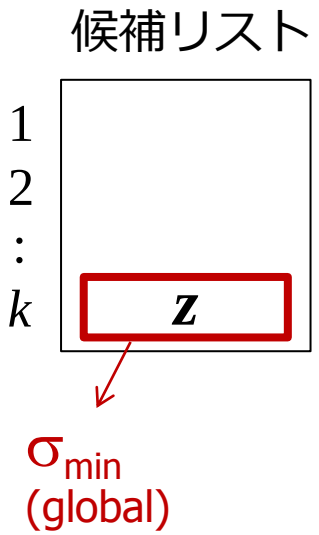
- 条件 $p(\neg c, x) = 0$ は以下と同値：
$p(c \mid x) = 1$, $p(x) = p(c)p(x \mid c)$, …

$p(x \mid c)$ の関数として表せる
（上の分割表の自由度は2）

「弱い」関連パターンの枝刈り (2)

- 接尾探索木において先祖・子孫は包含関係にある
  ➔ 包含検査をスキップできる
- 根から葉までのパスごとに局所的な最小サポートを導入し，パスを下るごとに最小サポートを上昇させる



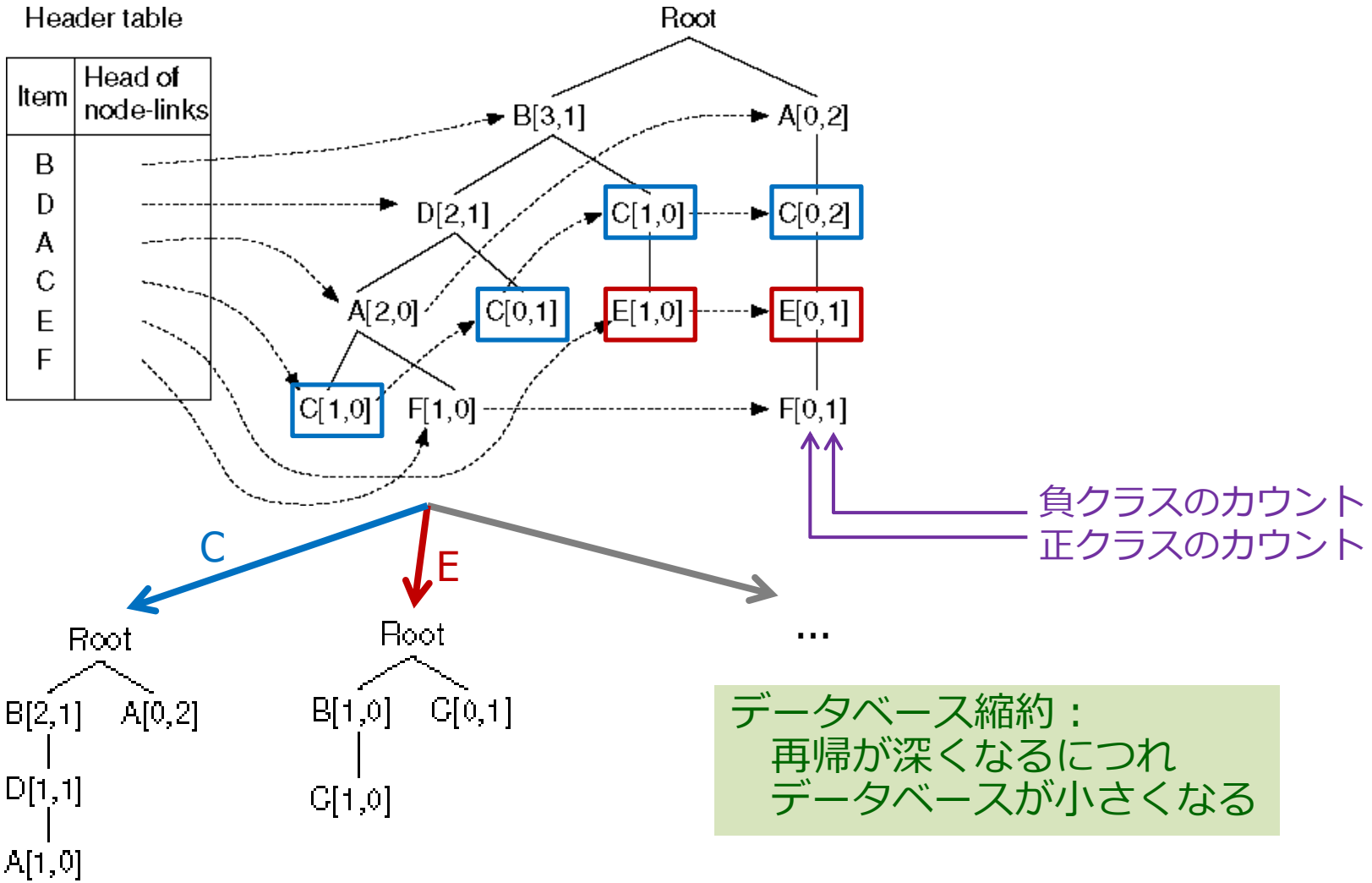


RP-tree

- FP-tree の拡張

Class c	Transaction
+	{A, B, C, D}
+	{A, B, D, F}
+	{B, C, E}
−	{A, C}
−	{B, C, D}
−	{A, C, E, F}

Item x	$N(+, x)$	$N(-, x)$	$F_+(x)$
B	3	1	0.857
D	2	1	0.667
A	2	2	0.571
C	2	3	0.500
E	1	1	0.400
F	1	1	0.400

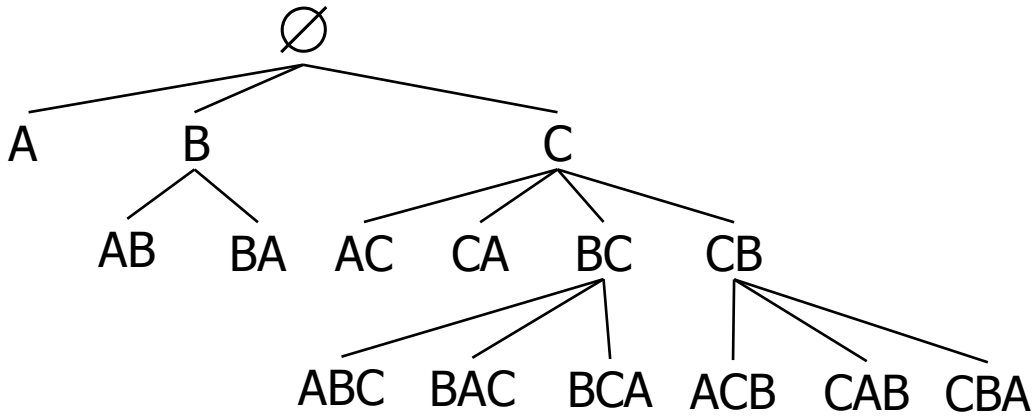


関連研究：系列データへの拡張

- RP-growth の特長
  - 凹性を満たさない関連スコアに対する分岐限定法
  - 最小サポート上昇への翻訳（データベース縮約）
  - 「弱い」関連パターンと枝刈り（接尾探索木の導入）

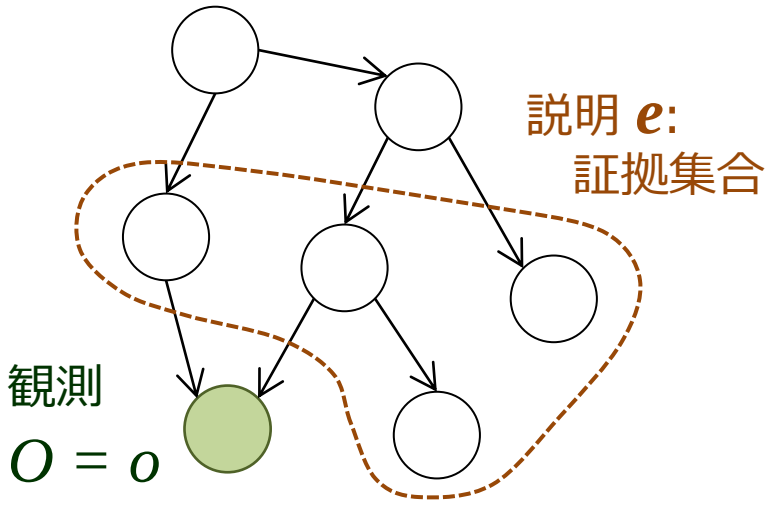
- これらの特徴は基本的に系列データにも適用可能
  - ただし接尾探索木におけるデータベースの射影操作は複雑になると思われる

例：順列（アイテムの重複のない系列）に対する接尾探索木



関連研究：ベイジアンネットの説明的分析

- 90年代初めに開発 (e.g. [Jensen 96])
- 説明 $\boldsymbol{e}$ と観測 o の依存関係を分析
- 典型的な関連スコア:
  [Chajewska & Halpern 97][Fishelson 07]
  - $p(\boldsymbol{e} \mid o)$
  - $p(o \mid \boldsymbol{e})$
  - …



- Most Relevant Explanation (MRE) [Yuan et al. 08 & 09]
  - Generalized Bayes factor (GBF) と呼ばれる関連スコアを使用
  - 観測 o に対し，説明 $\boldsymbol{e}$ を求める
    - MCMC に基づく探索
    - 安全でない枝刈りを用いた束上の探索

関連研究：「弱い」関連パターン

- パターン x' は x より弱い ⇔ $x \subset x'$ かつ $R_c(x) \geq R_c(x')$
- パターン x は非弱 ⇔ x は他のパターン x' と比べて弱くない
- 弱いパターンは冗長であるとして取り除く

- 既存研究でも類似の概念が導入されてきた：
  - 関係「より弱い」はベイジアンネットの説明的解析 [Yuan et al. 09] で導入された “strong dominance” の逆
  - 性質「非弱である」は，関連スコアを確信度（精度）としたときの性質 “productive” [Bayardo et al. 00] と同値

RP-growth: 分岐限定法の考えに基づき，
「弱い」関連パターンを枝刈りする点が新しい