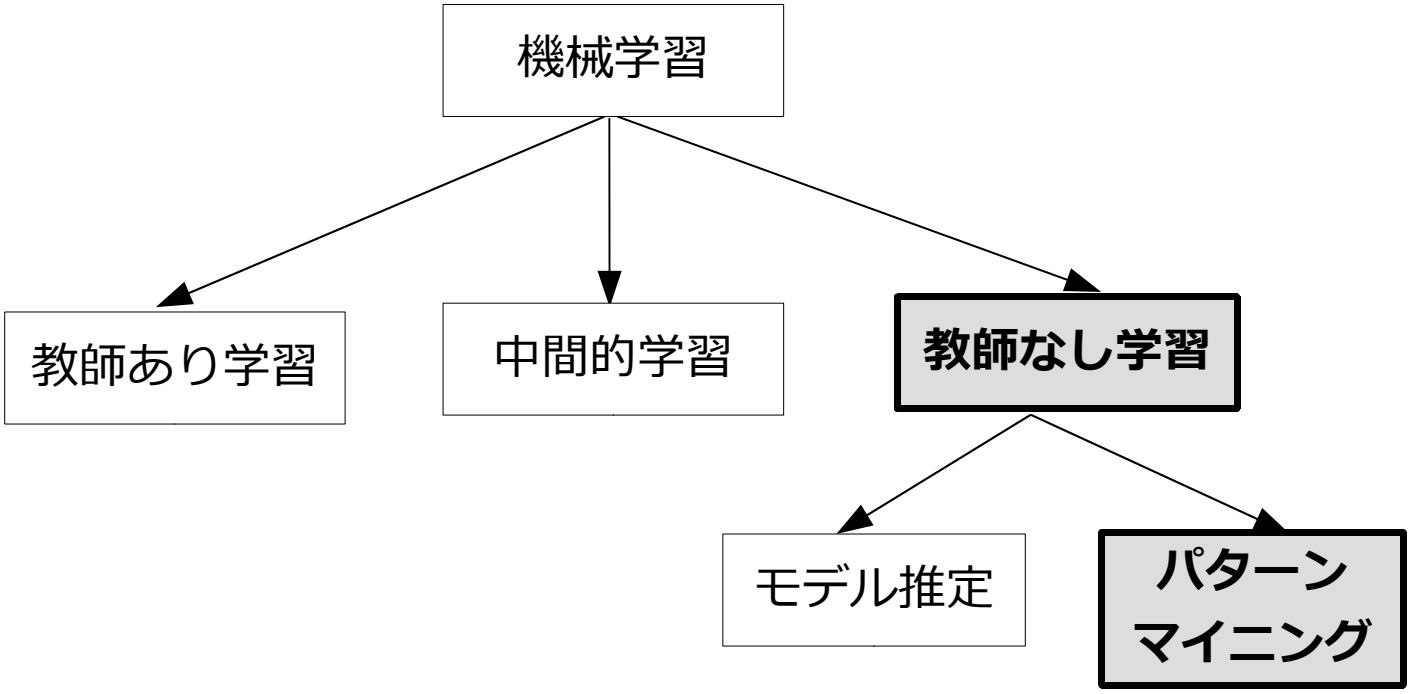
12. パターンマイニング

- パターンマイニングの問題設定
  - 入力：カテゴリ特徴の教師なしデータ
  - 出力：頻出項目、連想規則、未観測データ



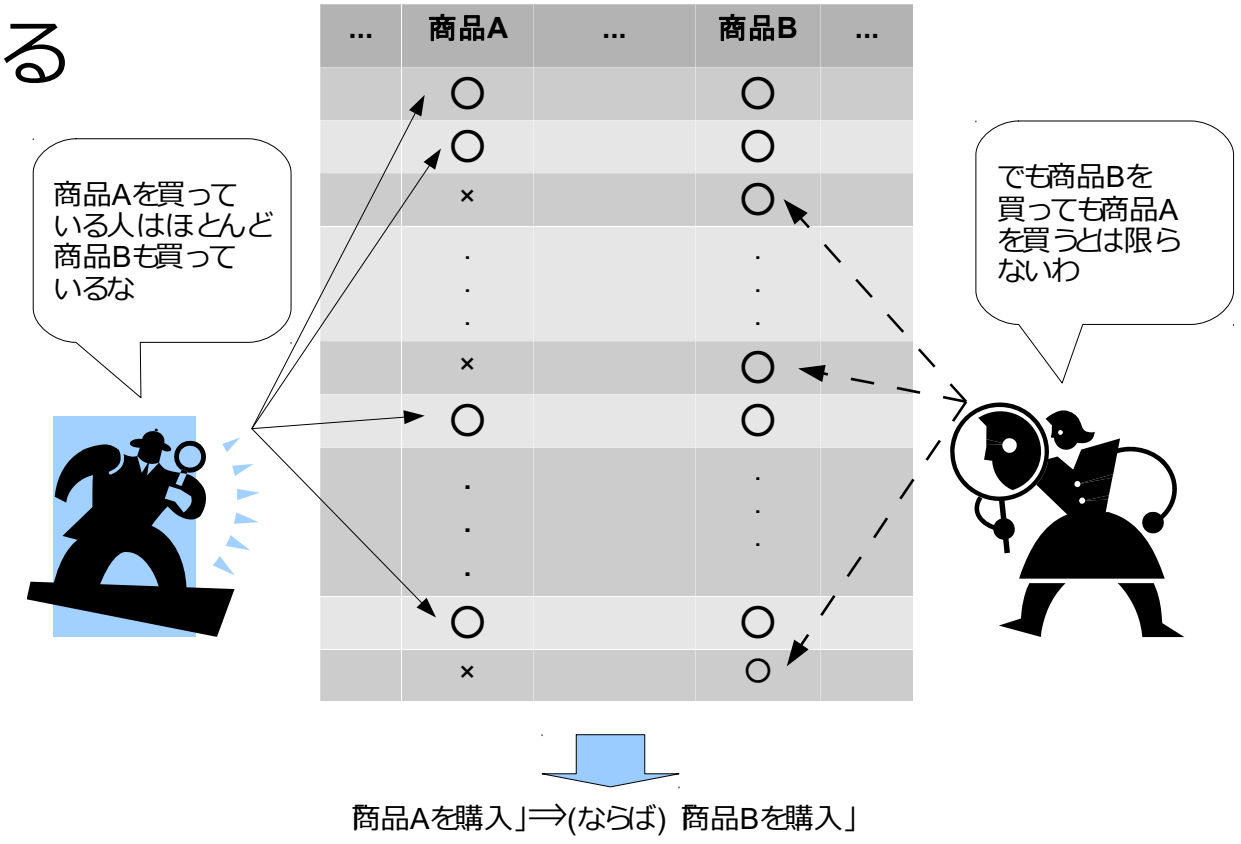
No.	ミルク	パン	バター	雑誌
1	t	t		
2		t		
3				t
4		t	t	
5	t	t	t	
6	t	t		

12.1 カテゴリ特徴に対する「教師なし・パターンマイニング」問題の定義

- 学習データ

$$\{\mathbf{x}^{(i)}\} \quad i = 1, .., N$$

- 問題設定
  - データ集合中で、一定頻度以上で現れるパターンを抽出する



12.2 頻出項目抽出

- 例題：バスケット分析

No.	ミルク	パン	バター	雑誌
1	t	t		
2		t		
3				t
4		t	t	
5	t	t	t	
6	t	t		

バスケット分析では、 1 件分のデータをトランザクションとよぶ

- 支持度
  - 全トランザクション数 T に対して、ある項目集合 (items) が出現するトランザクションの割合

$$support(items) = \frac{T_{items}}{T}$$

12.2.1 頻出の基準と問題の難しさ

- バスケット分析の目的
  - 支持度の値が閾値以上の項目集合を抽出したい
- バスケット分析の問題点
  - すべての可能な項目集合について、支持度を計算することは現実的には不可能

項目集合の種類数は 2 の商品数乗

商品数 1,000 の店なら 2^{1000}

高頻度の項目集合だけに絞って計算を行う必要がある

12.2.2 Apriori アルゴリズムによる頻出項目抽出

- a priori な原理

ある項目集合が頻出ならば、その部分集合も頻出である

例）「パン・ミルク」が頻出ならば「パン」も頻出

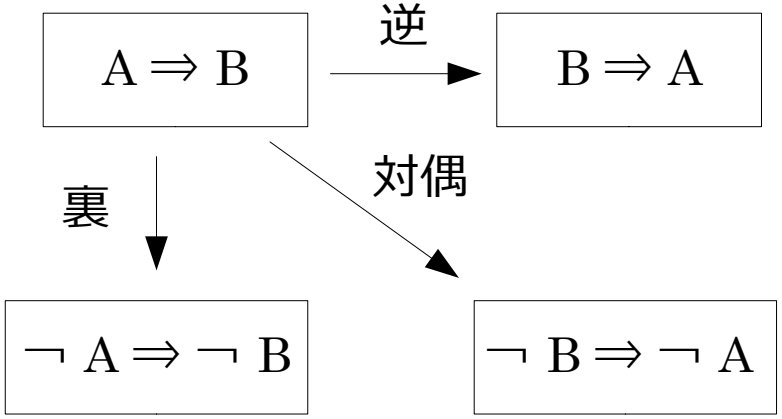
対偶

ある項目集合が頻出でないならば、
その項目集合を含む集合も頻出でない

例）「バター・雑誌」が頻出でないならば「バター・雑誌・パン」も頻出でない

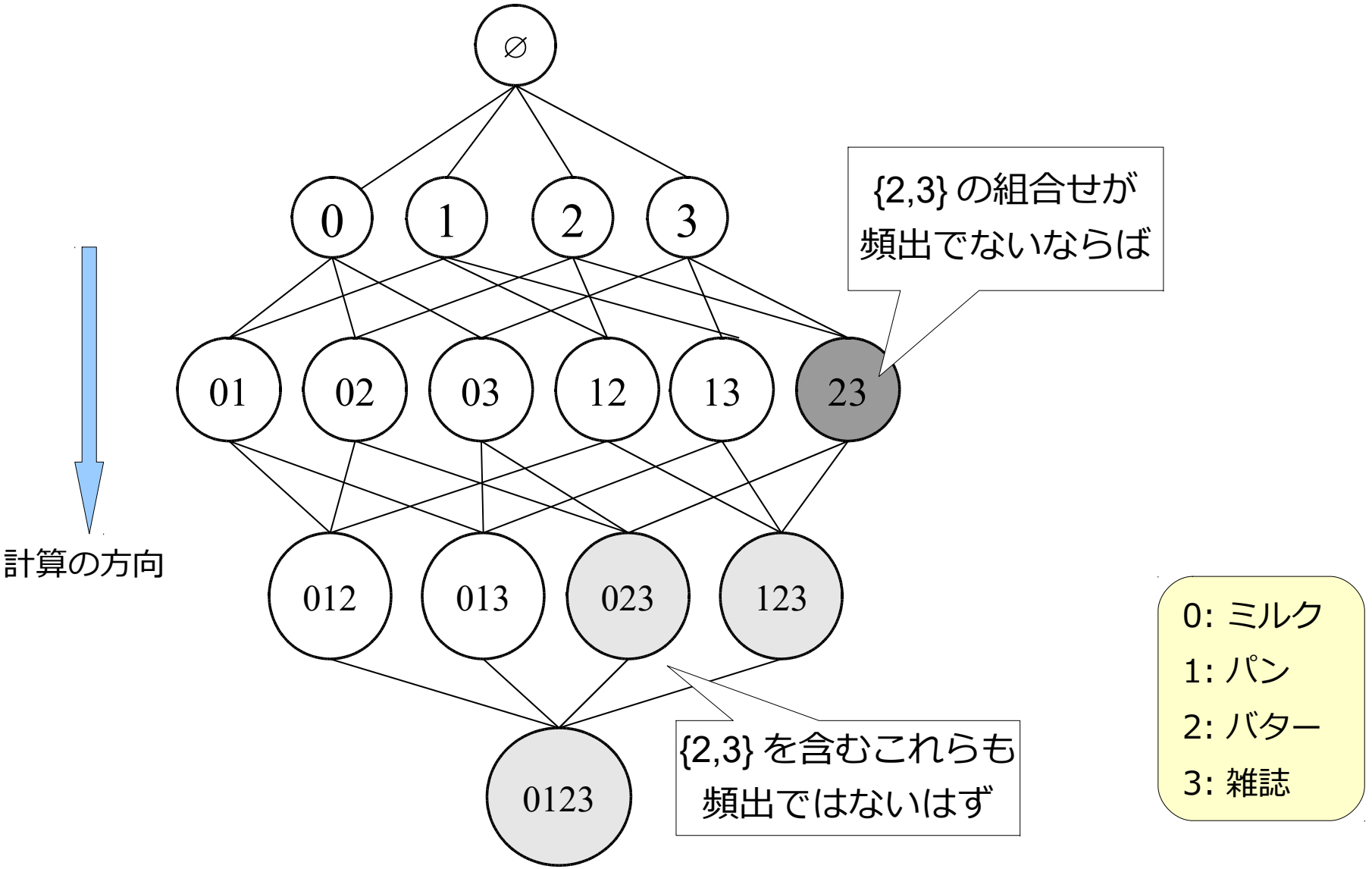
12.2.2 Apriori アルゴリズムによる頻出項目抽出

- 命題論理
  - 「 A ならば B 」が成り立つなら、必ずその対偶である「¬ B ならば¬ A 」が成り立つ



「 A ⇒ B 」は「¬ A∨B 」と定義されている。
一方、「¬ B ⇒ ¬ A 」は
　¬（¬ B ）∨ (¬ A)
なので、
　B ∨ ¬ A
となり、「¬ A∨B 」と等しい

12.2.2 Apriori アルゴリズムによる頻出項目抽出



12.2.2 Apriori アルゴリズムによる頻出項目抽出

Algorithm 12.1 Apriori アルゴリズム（頻出項目抽出）

入力: 正解なしデータ D

出力: 頻出項目集合

$F_1 \leftarrow$ 要素数 1 の頻出項目集合
$k = 2$
while $F_{k-1} \neq \emptyset$ **do**
  $C_k \leftarrow F_{k-1}$の各要素を組み合わせ
  for all $\boldsymbol{x} \in D$ **do**
    for all $c \in C_k$ **do**
      if $c \subset \boldsymbol{x}$ **then**
        $c.count \leftarrow c.count$
      end if
    end for
    $F_k \leftarrow \{c \in C_k \mid c.count >$ 閾値 $\}$
  end for
  $k \leftarrow k + 1$
end while
return $\bigcup_k F_k$

12.3 連想規則抽出

- 連想規則抽出の目的
  - 「商品 A を買った人は商品 B も買う傾向が強い」というような規則性を抽出したい
  - 確信度またはリフト値の高い規則を抽出

$$\mathrm{confidence}(\mathrm{A} \rightarrow B) = \frac{support(A \cup B)}{support(A)} = \frac{T_{A \cup B}}{T_A}$$

条件部 A が起こったときに結論部 B が起こる割合

$$\mathrm{lift}(\mathrm{A} \rightarrow B) = \frac{confidence(A \rightarrow B)}{support(B)}$$

B だけが単独で起こる割合と A が起こったときに B が起こる割合との比

12.3 連想規則抽出

- 支持度・確信度・リフト値
  - 砂糖について卵の関連購買が以下の場合：
    - 支持度 20%　確信度 70%　リフト値　30.0
  - 「全体顧客の 20% が砂糖と卵を一緒に購入しており、砂糖購入者の 70% が砂糖と卵を一緒に購入している」ということになる。この時のリフト値　30.0 は、「顧客全体の中で卵をいきなり購入するよりも、砂糖を買って卵を買う確率が 30 倍大きい」という意味を表している。

https://kotobank.jp/word/ リフト値 -801685　（訳語改変）

12.3 連想規則抽出

- 連想規則抽出の手順
  - 頻出項目集合を求める
  - 項目集合を条件部、空集合を結論部とした規則を作成する
  - 条件部から結論部へ項目を 1 つずつ移動し、評価する

12.3 連想規則抽出

- a priori な原理

ある項目集合を結論部に持つ規則が頻出ならば、
その部分集合を結論部に持つ規則も頻出である

例）結論部が「パン・ミルク」の規則が頻出ならば、結論部が「パン」の規則も頻出である

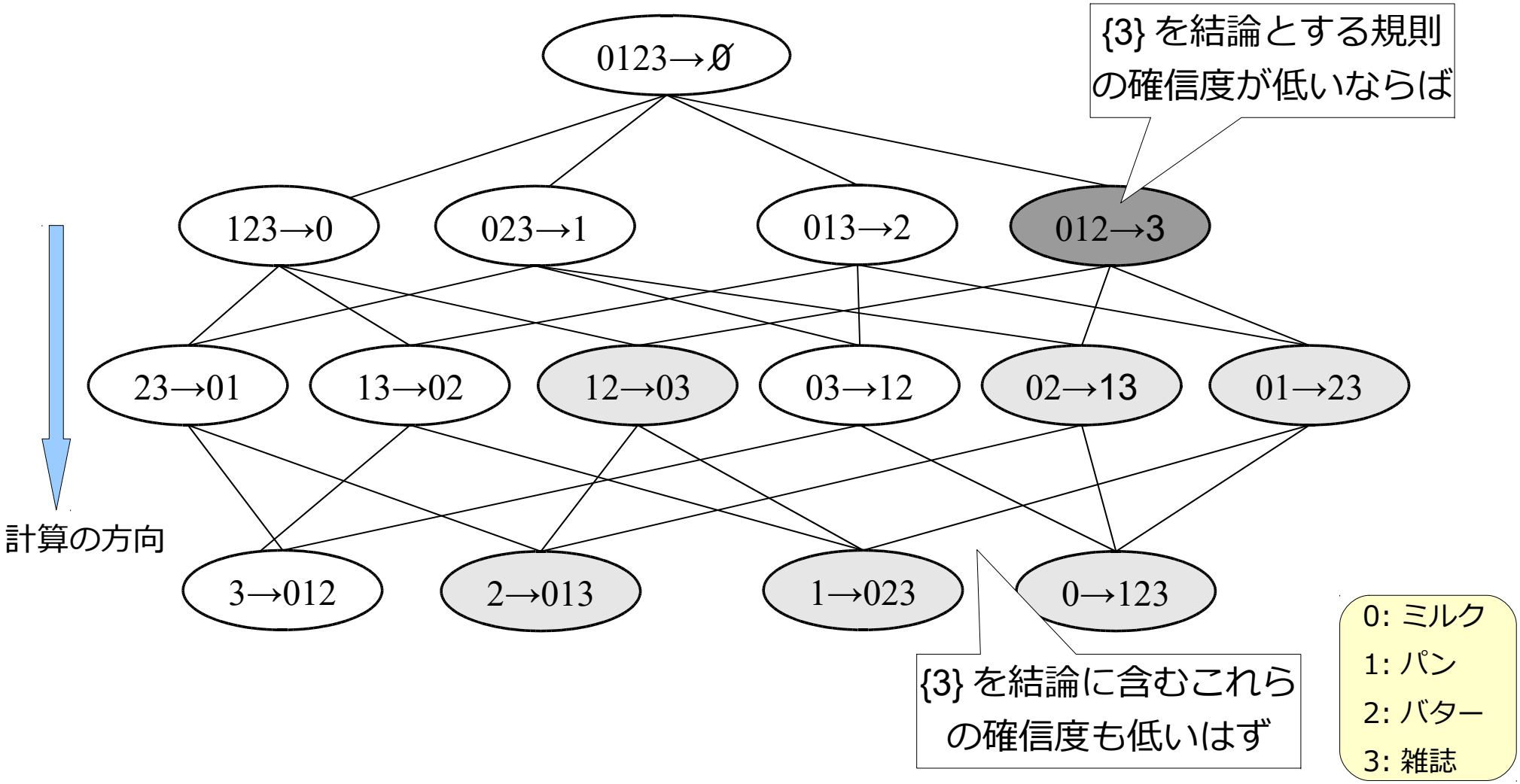
対偶

ある項目集合を結論部に持つ規則が頻出でないならば、
その項目集合を結論部に含む規則集合も頻出でない

例）結論部が「雑誌」の規則が頻出でないならば、結論部が「パン・雑誌」の規則も頻出でない

12.3 連想規則抽出

- a priori 原理に基づく探索



12.4 FP-Growth アルゴリズム

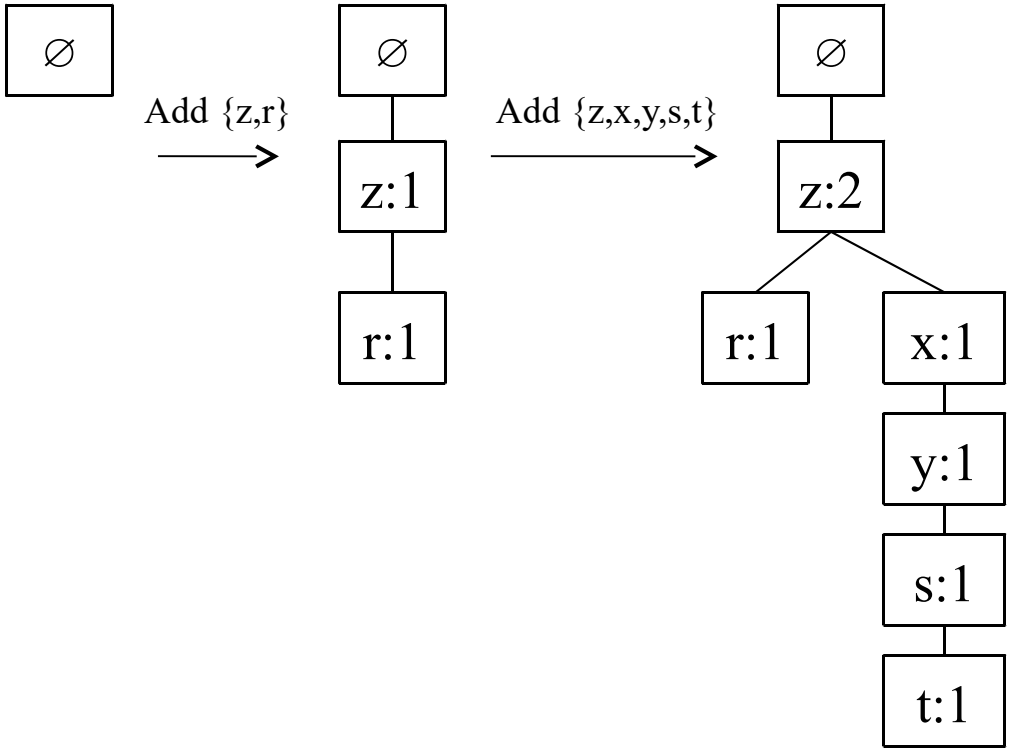
- Apriori アルゴリズムの高速化
  - トランザクションをコンパクトに表現し、重複計算を避ける
  - トランザクションの前処理
    - トランザクションを、出現する特徴名の集合に変換
    - 出現頻度順にソート
    - 低頻度特徴をフィルタリング

```
1 {r,z,h,j,p}
2 {z,y,x,w,v,u,t,s}
3 {z}
4 {r,x,n,o,s}
5 {y,r,x,z,q,t,p}
6 {y,z,x,e,q,s,t,m}
```

```
1 {z,r}
2 {z,x,y,s,t}
3 {z}
4 {x,s,r}
5 {z,x,y,r,t}
6 {z,x,y,s,t}
```

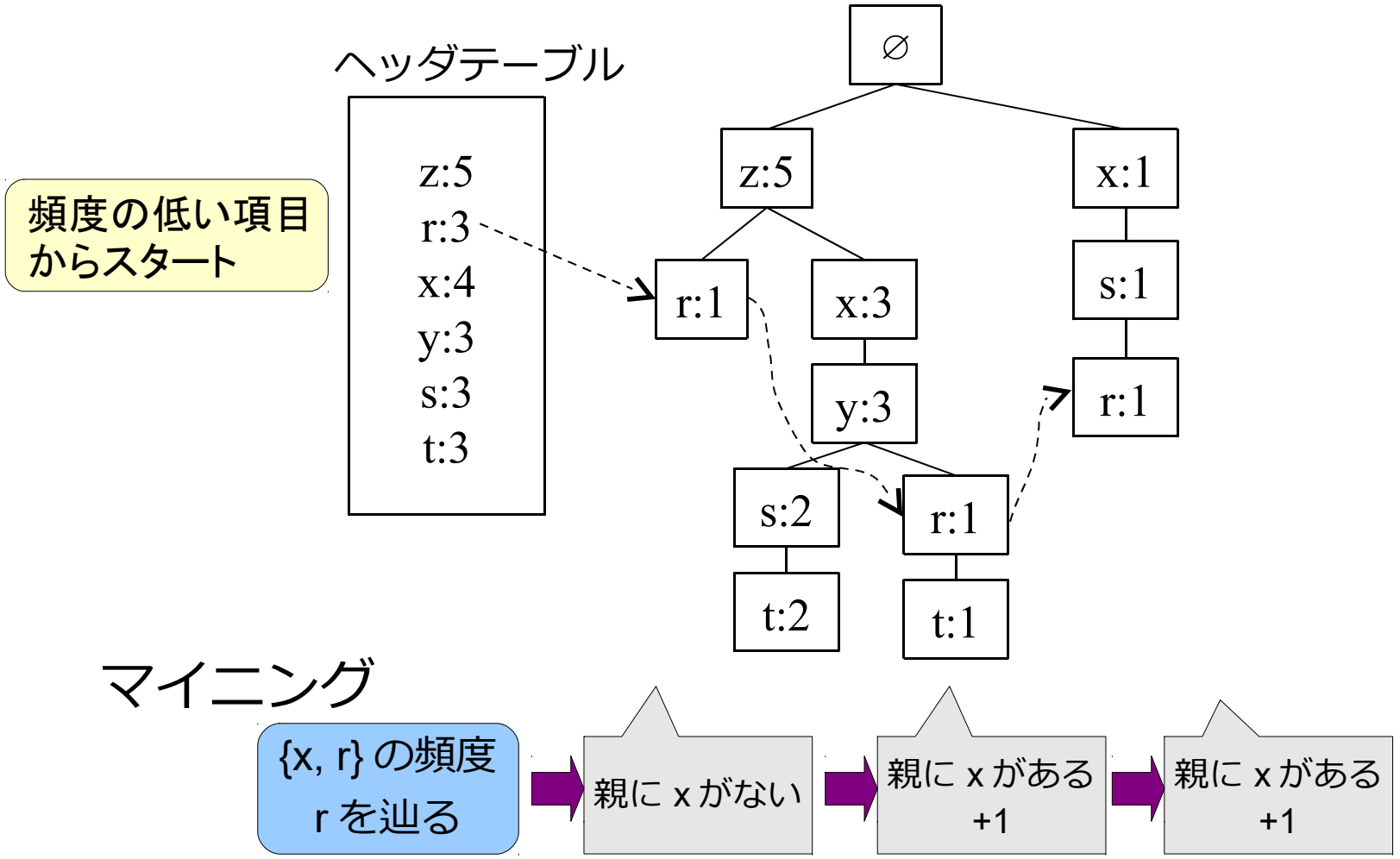
12.4 FP-Growth アルゴリズム

- トランザクションの表現 (FP 木)
  - ソート、フィルタリング後のトランザクションデータを順次 FP 木に挿入



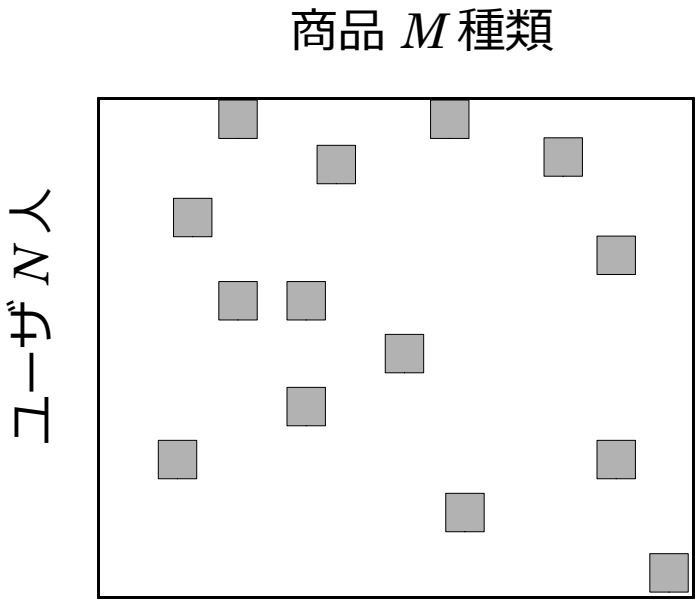
12.4 FP-Growth アルゴリズム

- FP 木のマイニング



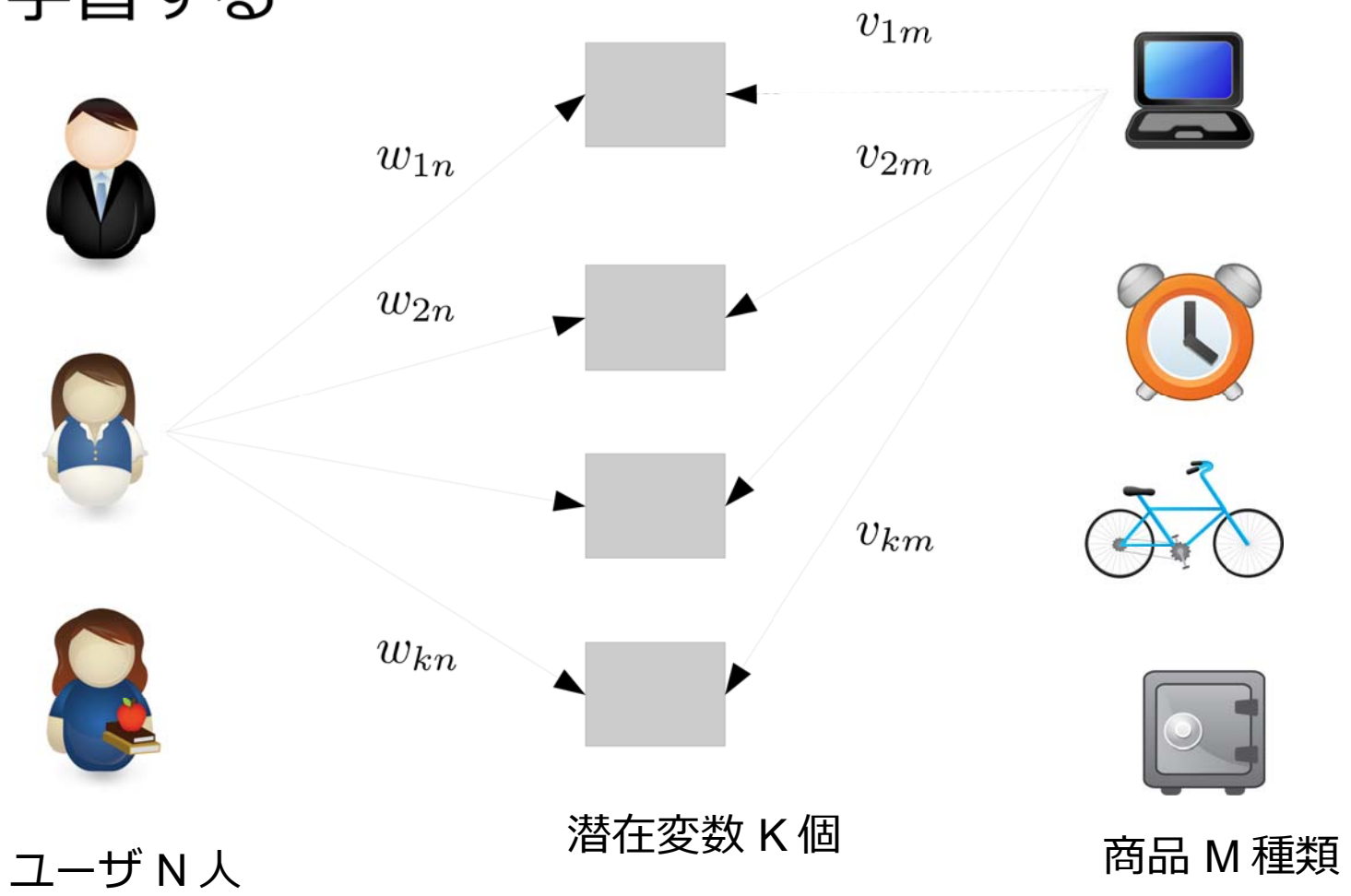
12.5 推薦システムにおける学習

- 問題設定
  - ある個人が、どの商品を購入したかという履歴がある
  - その個人が購入しそうな未購入商品を推薦したい



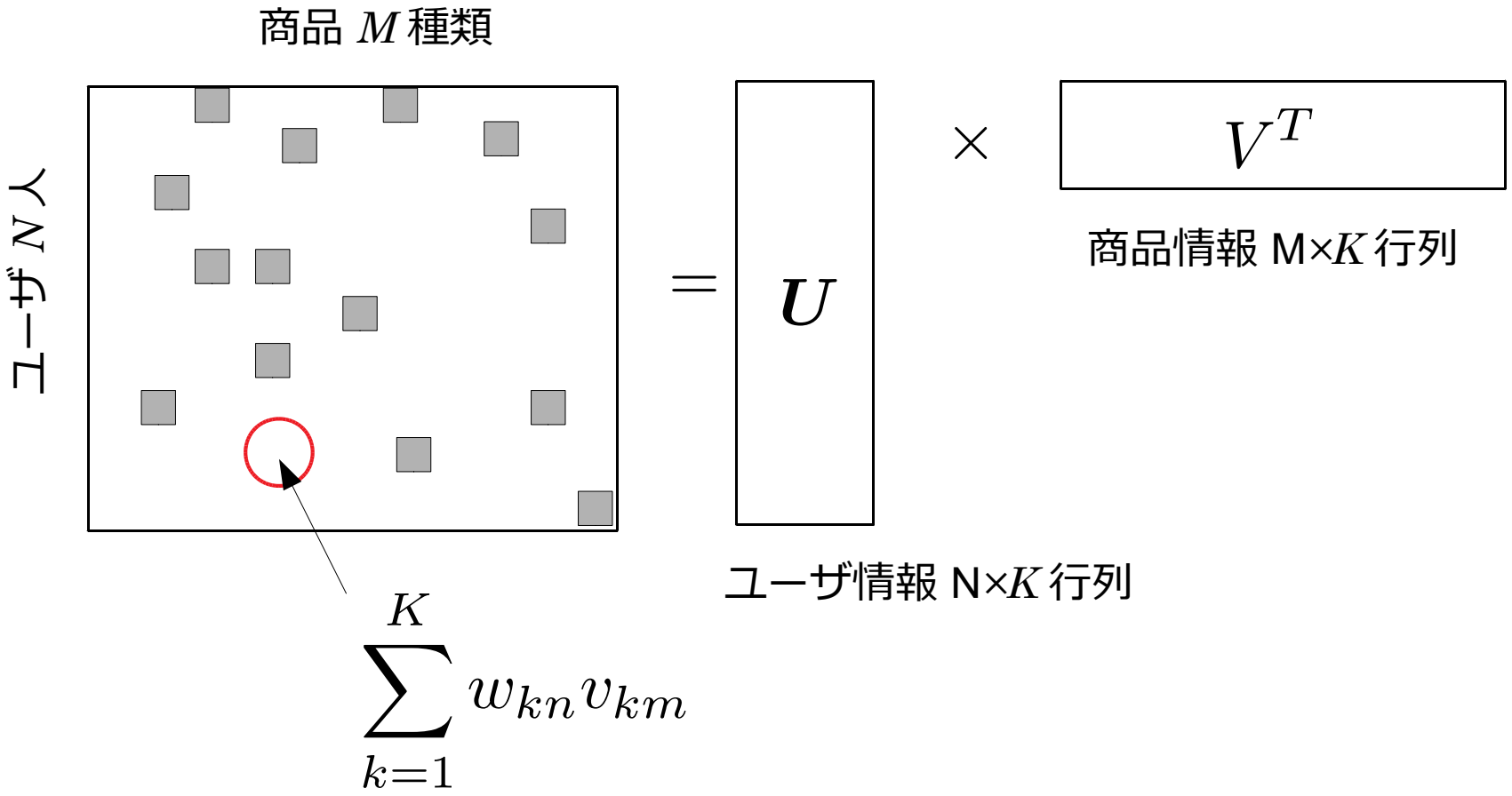
12.5 推薦システムにおける学習

- アプローチ
  - 個人、商品のそれぞれと潜在変数との関係を教師なしで学習する



12.5 推薦システムにおける学習

- 行列分解による潜在変数の抽出



12.5 推薦システムにおける学習

- 行列分解の方法
  - 誤差の最小化としての定式化

$$\min_{\boldsymbol{U},\boldsymbol{V}} \frac{1}{2}\|\boldsymbol{E}\|_{\mathrm{Fro}}^2 = \min_{\boldsymbol{U},\boldsymbol{V}} \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^T\|_{\mathrm{Fro}}^2$$

    - 値がないところを 0 と解釈している
  - Alternating Least Squares 法

$$\min_{\boldsymbol{U},\boldsymbol{V}} \sum_{(i,j)\in\Omega} (x_{ij} - \boldsymbol{u}_i^T \boldsymbol{v}_j)^2 + \lambda_1 \|\boldsymbol{U}\|_{\mathrm{Fro}}^2 + \lambda_2 \|\boldsymbol{V}\|_{\mathrm{Fro}}^2$$

    - 値があるところだけで誤差最小化＋正則化