参加システムの嗜好パターンが異なる場合の集団協調フィルタリング

神嶌 敏弘 (http://www.kamishima.net), 赤穂 昭太郎
産業技術総合研究所

人工知能学会 FPAI研究会 (2007/11/3)

概要

協調フィルタリング

- 利用者が少ないとうまくいかない

集団協調フィルタリング

- 複数サイトを集めて利用者数を増やす
  - 広域ネットワーク上に分散 → 通信量を抑制
  - 個人情報の保護 → 個人嗜好データは局所サイト内でのみ保持
  - 各サイトに適応させた推薦モデルの獲得

実現のアイデア

- 分布パラメータの次元縮約による方法
- 大域潜在変数を導入する方法

推薦システム

情報過多 Information Overload

膨大な情報の集積

社会の高度情報化 & 情報発信の低コスト化
記憶媒体の大容量化 & 通信の高速化

↓

情報があっても利用できない

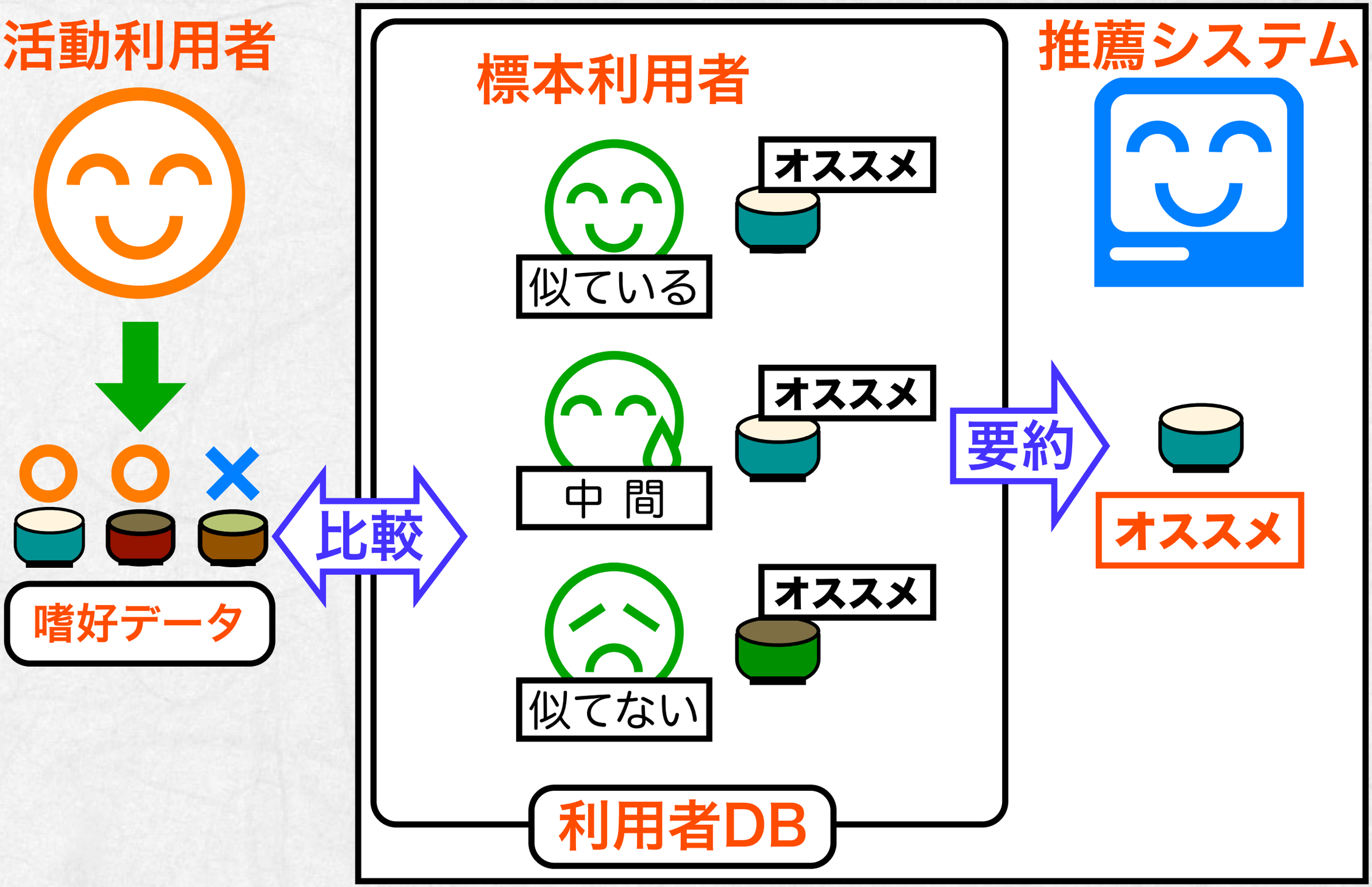
欲しい情報が埋もれている
必要な情報を具体化できない

↓

推薦システム Recommender System

利用者が必要としていると思われる情報を選び出す

協調フィルタリング

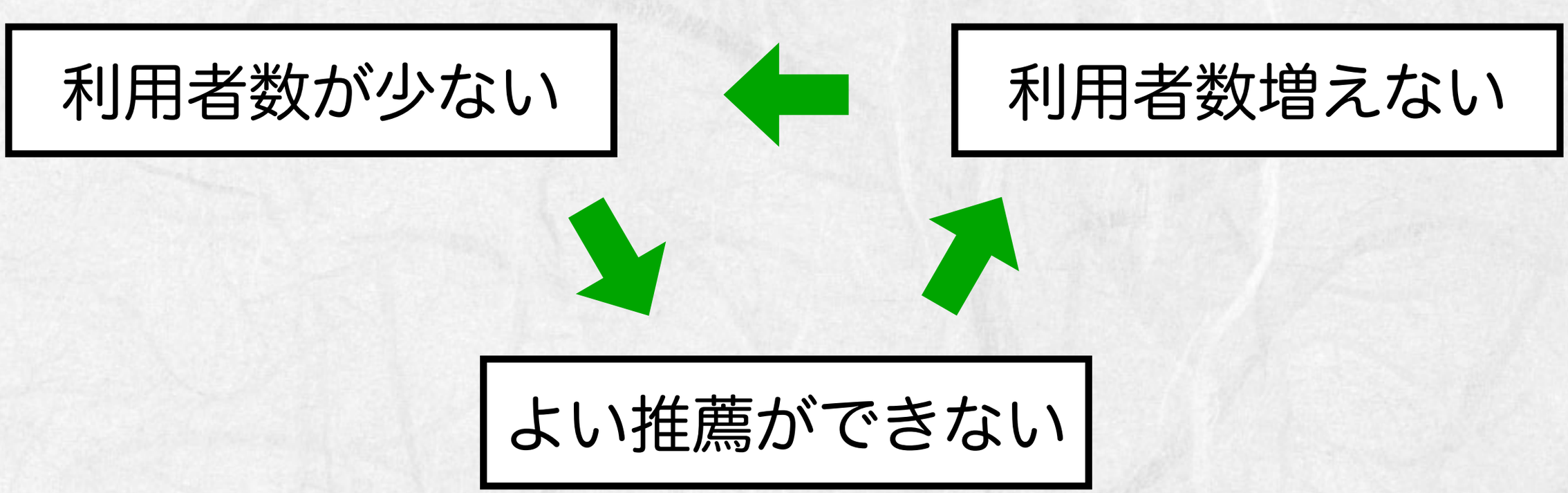

活動利用者
標本利用者
推薦システム
オススメ
似ている
中 間
オススメ
要約
オススメ
比較
オススメ
似てない
嗜好データ
利用者DB

協調フィルタリングと利用者数

協調フィルタリングは利用者数が少ないと稼働しない

協調フィルタリングシステムの運用を始めると…

負のフィードバック・ループ



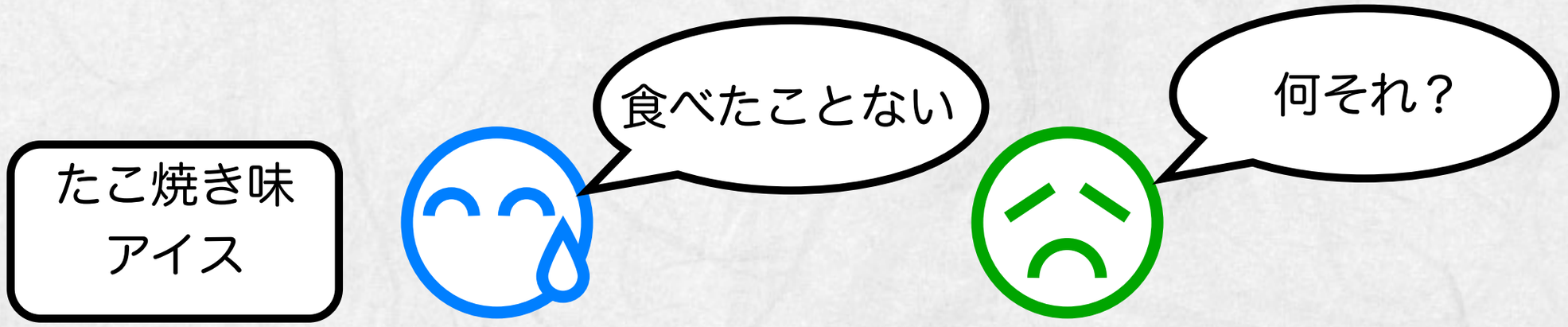
問題点は主に二つ

被覆率

データベース中には登録されていても
誰にも評価されていないアイテムは推薦されない

$$\text{被覆率(Coverage)} = \frac{\text{推薦候補にできるアイテム数}}{\text{とりそろえた全アイテム数}}$$

利用者数 少ない → 未評価アイテム 多い → 被覆率 低い



少数派の利用者

嗜好パターンが少数派の利用者には
嗜好パターンが類似している標本利用者がいない

利用者数
少ない

少数派の
嗜好

類似利用者
いない

嗜好パターン
無視

利用者数を増やすには

評価付けにインセンティブ

- サイトで商品が買える商品券や現金を配布する
- システムの利用にポイントが必要で，評価付けでポイントを配布 [Melamed 2007]

複数のシステムを統合

分散協調フィルタリング

複数の計算機を使って協調フィルタリングを実行

目的：計算の高速化 & プライバシーの確保

集団協調フィルタリング

利用者数を増やす目的で，複数の協調フィルタリングシステムのデータをまとめて扱う

サイト適応型集団協調フィルタリング

プライバシーの保持

- 個人嗜好データは，それを復元できない形で中心サイトに送る
- 要約情報（確率モデルの十分統計量）だけを送信する

疎な分散環境

- データはクラスタ環境のようなLANではなく，各サイトは広域ネットワークで接続されている
- 中心サイトに，小規模のデータを集めて計算

参加サイトの個性の考慮

- 参加しているサイトの利用者グループには，独自の特徴があり，一つの推薦モデルではそれらに十分には対応できない
- 個別のサイトに適応させたモデルを構築

pLSAによる協調フィルタリング

pLSAによる手法を拡張して
サイト適応型集団協調フィルタリングを実現

pLSA (probabilistic Latent Semantic Analysis)

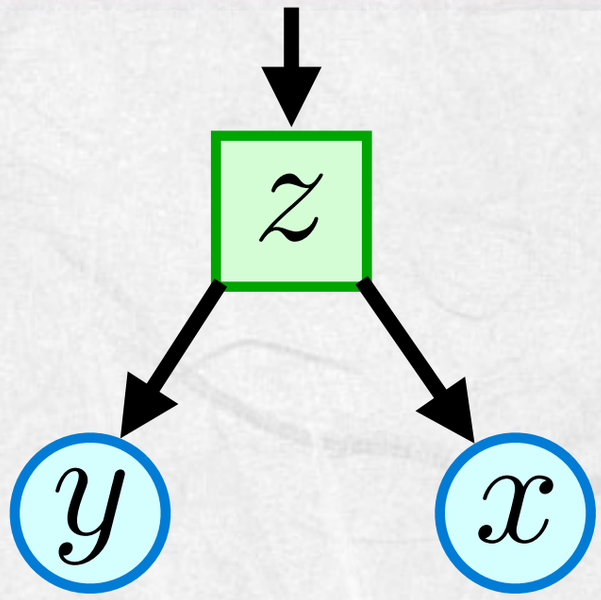
[Hoffmann 1999]による自然言語処理のための次元縮約法

- [Hoffman & Puzicha 1999]で協調フィルタリングにも適用
- 未評価と否定的評価の区別がなくても適用しやすい
- モデルの複雑さが利用者数やアイテム数に対して線形にしか増加しない
- 並列計算が容易

pLSA (1)

pLSAモデル

利用者： $x \in \{1, \ldots, n\}$
アイテム： $y \in \{1, \ldots, m\}$
潜在変数： $z \in \{1, \ldots, l\}$



生成モデル： z が与えられたとき x や y は条件付独立

同時分布： $\Pr[x, y] = \sum_{z \in Z} \Pr[x|z] \Pr[y|z] \Pr[z]$

訓練データ： $\mathcal{D} = \{(x = i_k, y = j_k)\}_{k=1}^{N}$

アイテム j_k を利用者 i_k が購入/閲覧した

尤度関数： $\mathcal{L}(\mathcal{D}; \theta) = \sum_{(i,j) \in \mathcal{D}} \ln \Pr[x = i, y = j; \theta]$

pLSA (2)

pLSAのEMアルゴリズムによる最尤推定

パラメータ： $\theta = (\{\Pr[x|z]\}, \{\Pr[y|z]\}, \{\Pr[z]\})$

ステップ1：与えられたパラメータから潜在変数を計算

$$\Pr[z|x,y] = \frac{\Pr[z]\Pr[x|z]\Pr[y|z]}{\sum_{z'}\Pr[z']\Pr[x|z']\Pr[y|z']}$$

ステップ2：与えられた潜在変数からパラメータを計算

$$\Pr[x|z] = \frac{\sum_{y} n(x,y)\Pr[z|x,y]}{\sum_{x',y} n(x',y)\Pr[z|x',y]}$$

データ対

$\Pr[y|z]$ や$\Pr[z]$についても同様に計算

推薦： x=i のときの事後確率を最大化するアイテム

$$y^* = \arg\max_{y\in\{1,\ldots,m\}} \Pr[y|x=i]$$

並列pLSA

利用者のデータを二つのサイトで分割して保持

サイト1	サイト2
$\mathcal{X}_1 = \{1, \ldots, n_1\}$	$\mathcal{X}_2 = \{n_1 + 1, \ldots, n\}$

サイト1での値のステップ2の計算

$$\Pr[x|z] = \frac{\sum_y n(x,y) \Pr[z|x,y]}{\sum_{x',y} n(x',y) \Pr[z|x',y]}$$

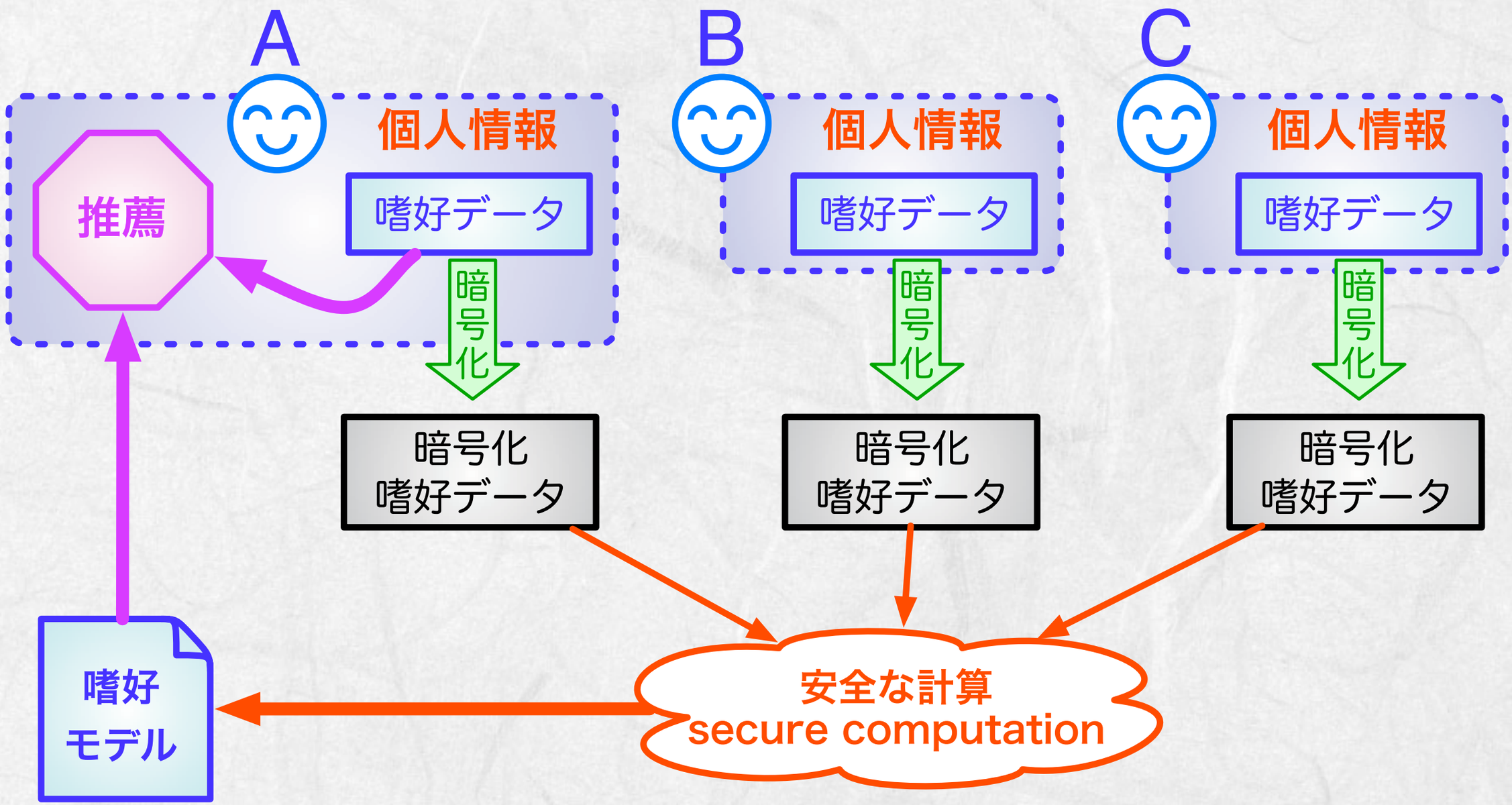
$x' \in \mathcal{X}_2$ の利用者については $n(x',y)$ の値は未知

$$\Pr[x|z] = \frac{\sum_y n(x,y) \Pr[z|x,y]}{\sum_{x' \in \mathcal{X}_1, y} n(x',y) \Pr[z|x',y] + \sum_{x' \in \mathcal{X}_2, y} n(x',y) \Pr[z|x',y]}$$

サイト2からサイト1に送信

プライバシー協調フィルタリング

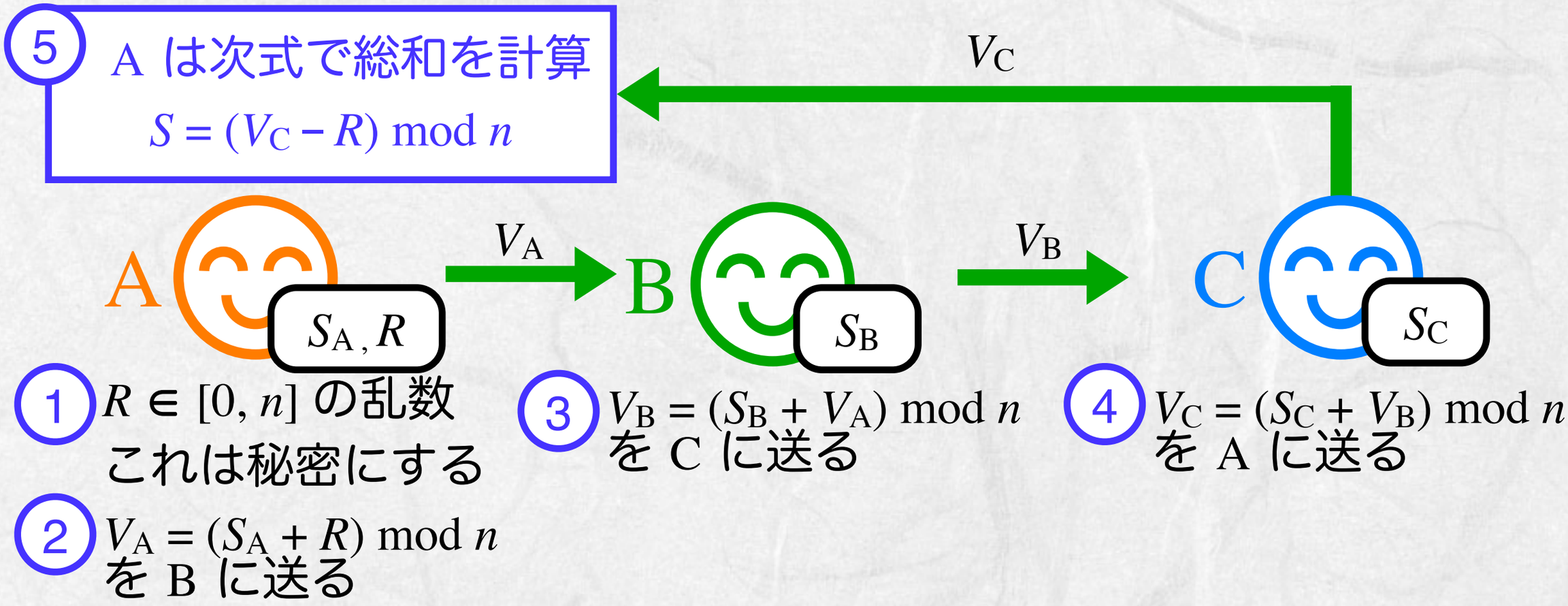
嗜好データは暗号化してから外部に出す



安全な計算

A B C の3人が，それぞれ，秘密の値 S_A S_B S_C を保持
S_A S_B S_C を他人に明かさず総和 $S = S_A + S_B + S_C$ を計算

n は S より大きな数



厳密な値が計算できるが，共謀には脆弱

pLSAとプライバシー保護

semi-honest

個人情報を明かすほどには信用はできないが，計算のプロトコルは順守することぐらいは信用できる

集団協調フィルタリングでは，参加サイトは団体

↓

社会的手段によって semi-honest を保証できると仮定

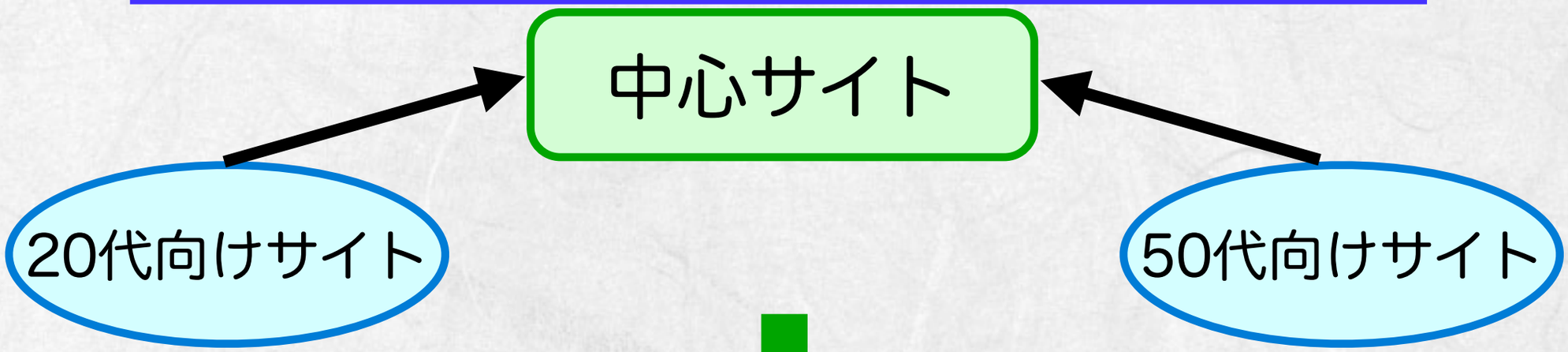
個人情報かどうか?

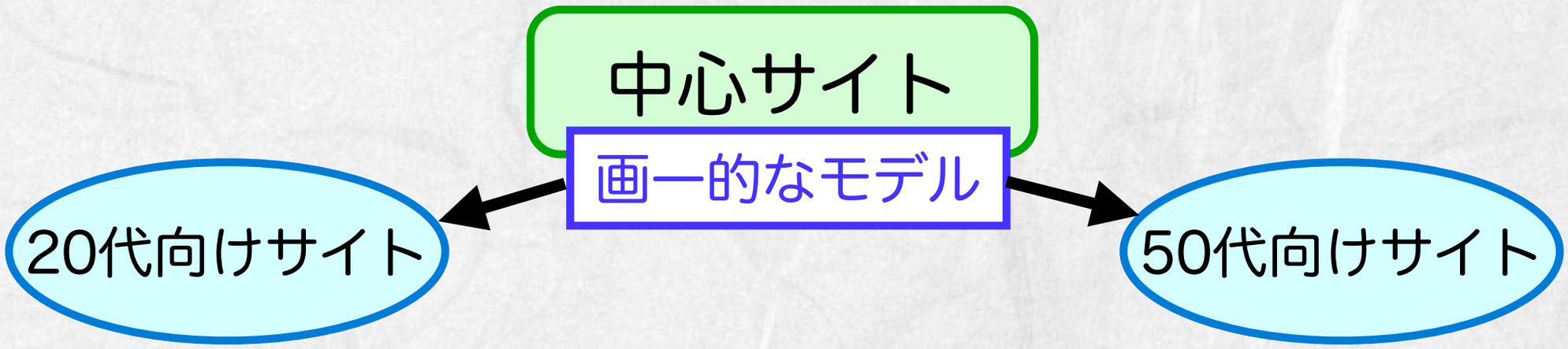
$\Pr[z]$ と $\Pr[y|z]$ は共に x とは無関係なので個人情報ではない

$\Pr[x|z]$は個人の嗜好パターンの記述 → **個人情報**

参加サイトの個性の考慮

個性の異なるサイトのデータを集める





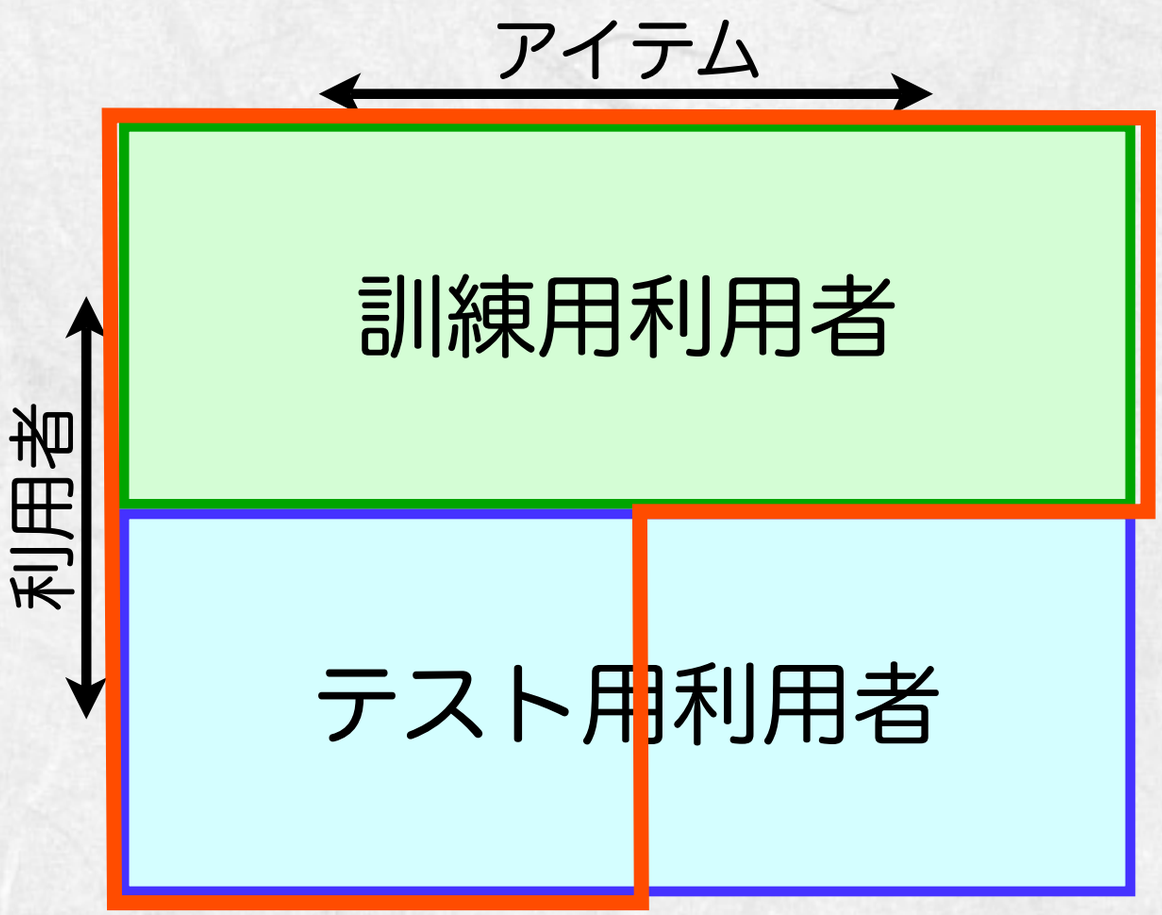
サイトの利用者の偏りが推薦に反映されない

サイトの個性：実験

映画の推薦MovieLensデータ

5段階のうち4か5の評価なら映画を肯定的に評価

- 利用者を訓練用とテスト用に分ける
- 訓練利用者の全ての評価とテスト用利用者の半分の評価を訓練データ(赤枠)
- pLSA (潜在変数 $l=10$) を適用し，テスト利用者が肯定的に評価したた残りのアイテムに割り当てられた確率質量の利用者ごとの総和の，全テスト利用者の総和
- 全アイテムを評価していて，予測が完全なら最大値 0.5



サイトの個性：結果

テスト集合	平均確率	人数
ベースライン	0.0695	189
20歳未満	0.0664	77
20歳代	0.0747	332
30歳代	0.0706	240
40歳代	0.0593	168
50歳以上	0.0610	125
60歳以上	0.0559	31

少数派集団への予測精度は悪い

広域分散環境下でのpLSA

並列pLSAを広域ネットで実行

- 個人の嗜好データは復元できないのでプライシーの問題はない

しかし！

EMアルゴリズムの各反復で統計量をbroadcast

$$\sum_{x' \in \mathcal{X}_K, y} n(x', y) \Pr[z|x', y]$$

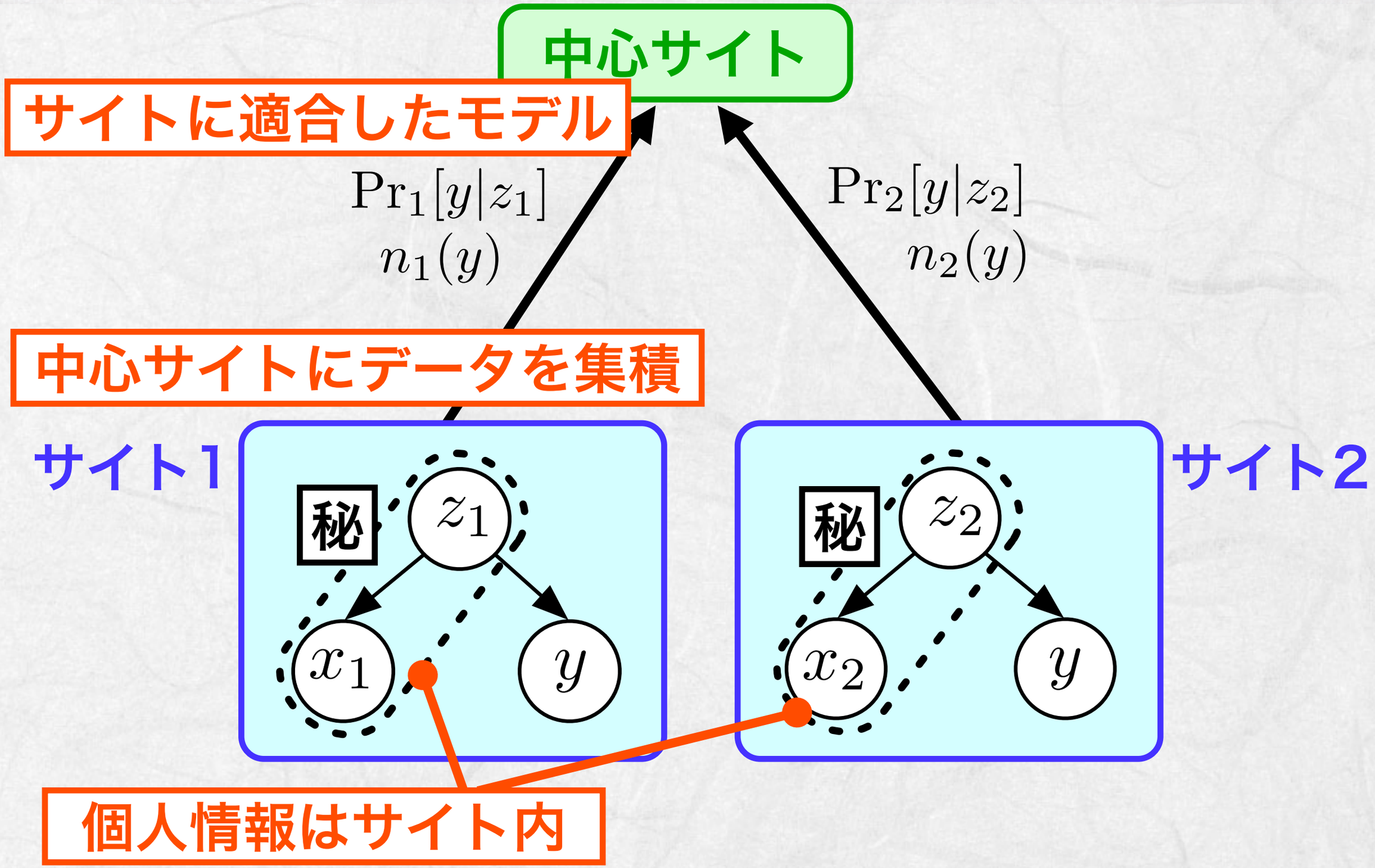
↓

- 広域ネットワークで各反復ごとに同期は難しい
- クラスタ構成のマシンより通信量の制約は強い

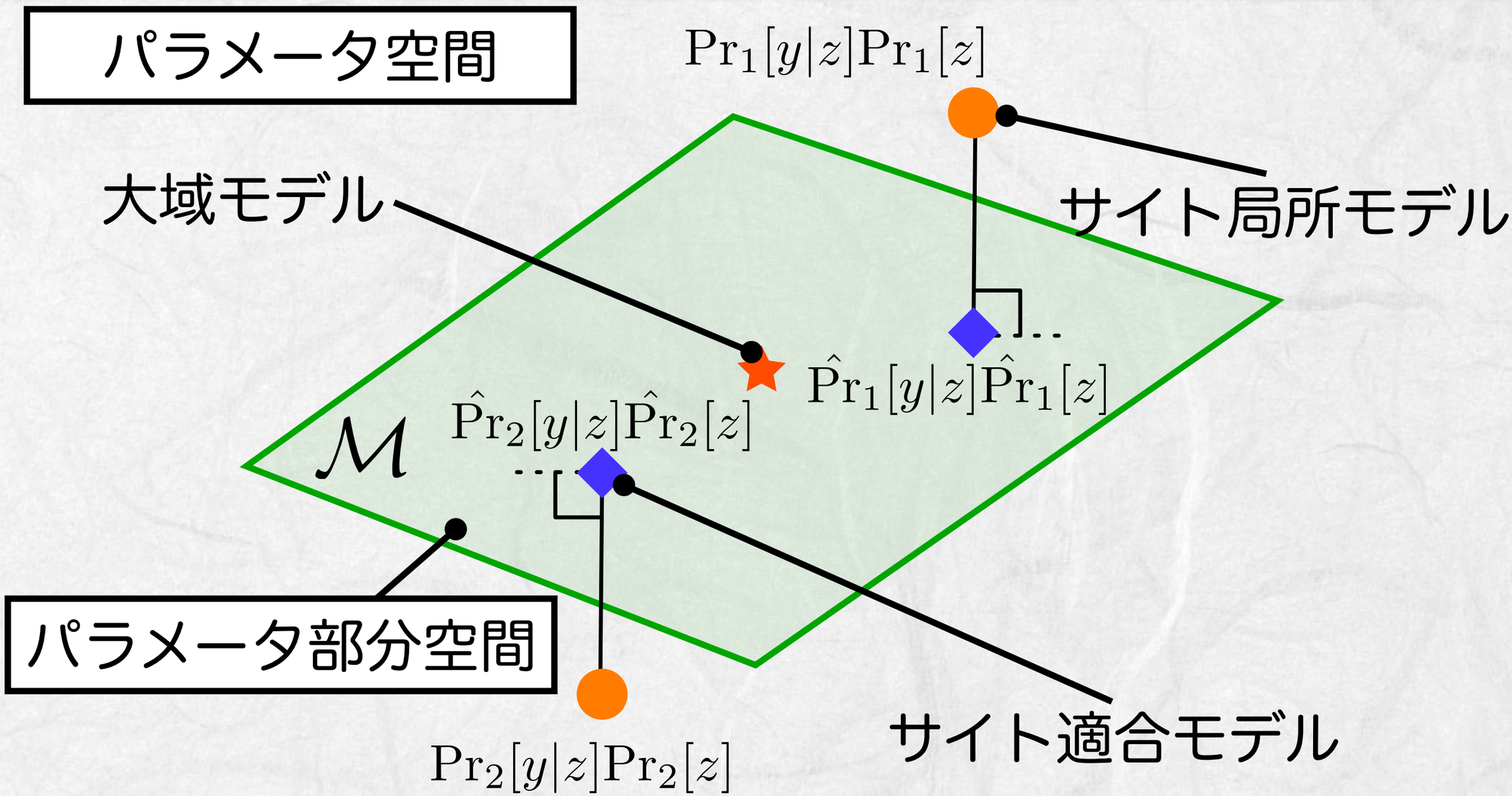
↓

データを中心サイトに集めて計算

サイト適応型集団協調フィルタリング

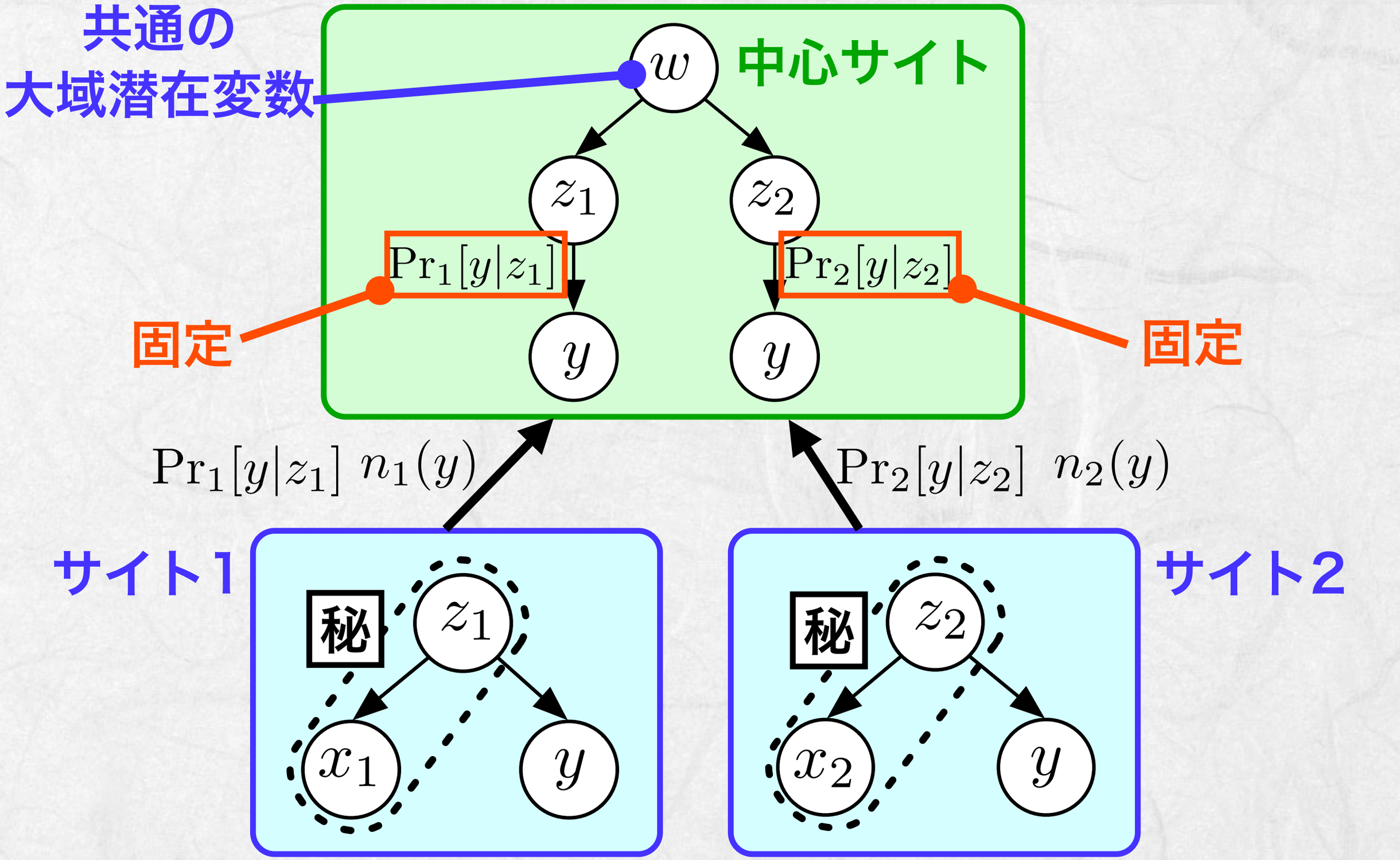


分布パラメータの次元縮約



情報幾何の考えに基づく距離や射影を利用

大域潜在変数の導入 (1)


共通の
大域潜在変数
w
中心サイト
z_1
z_2
$\Pr_1[y|z_1]$
$\Pr_2[y|z_2]$
固定
固定
y
y
$\Pr_1[y|z_1]$ $n_1(y)$
$\Pr_2[y|z_2]$ $n_2(y)$
サイト1
サイト2
秘
秘
z_1
z_2
x_1
x_2
y
y

大域潜在変数の導入 (2)

サイト1で評価されたアイテム

$$\Pr[y] = \sum_{z_1, z_2, w} \Pr_1[y|z_1] \Pr[z_1|w] \Pr[z_2|w] \Pr[w]$$

潜在変数が分かった場合の対数尤度

$$\log \mathcal{L}_1 = \sum_y \sum_{z_1, z_2, w} n_1(y) \Pr'[z_1, z_2, w|y] \log \left[\Pr_1[y|z_1] \Pr[z_1|w] \Pr[z_2|w] \Pr[w] \right]$$

$\log \mathcal{L}_1 + \log \mathcal{L}_2$ を最大化して大域パラメータを求める

局所サイトの事前分布を返す

$$\Pr^{new}[z_1] = \sum_{z_2, w} \Pr[z_1] \Pr[z_2|w] \Pr[w]$$

まとめ

サイト適応型集団協調フィルタリング

- **広域ネットワーク上に分散**
  - 各サイトで個別に学習したパラメータだけを中心サイトに送る
  - 計算が終わるまで，サイト間の通信は不要
- **個人情報の保護**
  - 個人の嗜好データである分布パラメータ $\Pr[x|z]$ は外部に送信せず，$\Pr[y|z]$ や $\Pr[z]$ だけを中心サイトに送る
- **各サイトに適応させた推薦モデルを獲得する**
  - 分布パラメータの次元縮約による方法
  - 大域潜在変数を導入する方法

今後の予定

- 提案したサイト適応型集団協調フィルタリング手法の実装・実験