転移学習を利用した
集団協調フィルタリング

神嶌 敏弘，赤穂 昭太郎
産業技術総合研究所
2009年度人工知能学会全国大会 (2009/6/17-19)
http://www.kamishima.net/

概要

協調フィルタリング

- **Start-up問題**：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

- 複数サイトの情報をマルチタスク学習を利用して集める
  - 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
  - 各サイトの個性の保持➡個別の推薦モデルの獲得

実現のアイデア

- 分散環境での階層モデルの提案
  - 実験結果は良くなかった

推薦システム

情報過多

膨大な情報の集積

↓

情報があるのに利用できない

欲しい情報が埋もれている or 必要な情報を具体化できない

推薦システム

利用者が必要としていると思われる情報を選び出す

内容ベース
フィルタリング
Content-Based Filtering

アイテムの特徴を利用

協調
フィルタリング
Collaborative Filtering

他人の意見を利用

協調フィルタリング

他人の意見を利用した「口コミ」による推薦

嗜好が似ている人が好きなものを推薦する

協調フィルタリングについてはあとで詳しく

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…

start-up問題：負のフィードバック・ループ

利用者数が少ない ← 利用者数増えない

↘ ↗

よい推薦ができない

- データベース中にはあっても，誰にも評価されていないアイテムは推薦されない
- 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

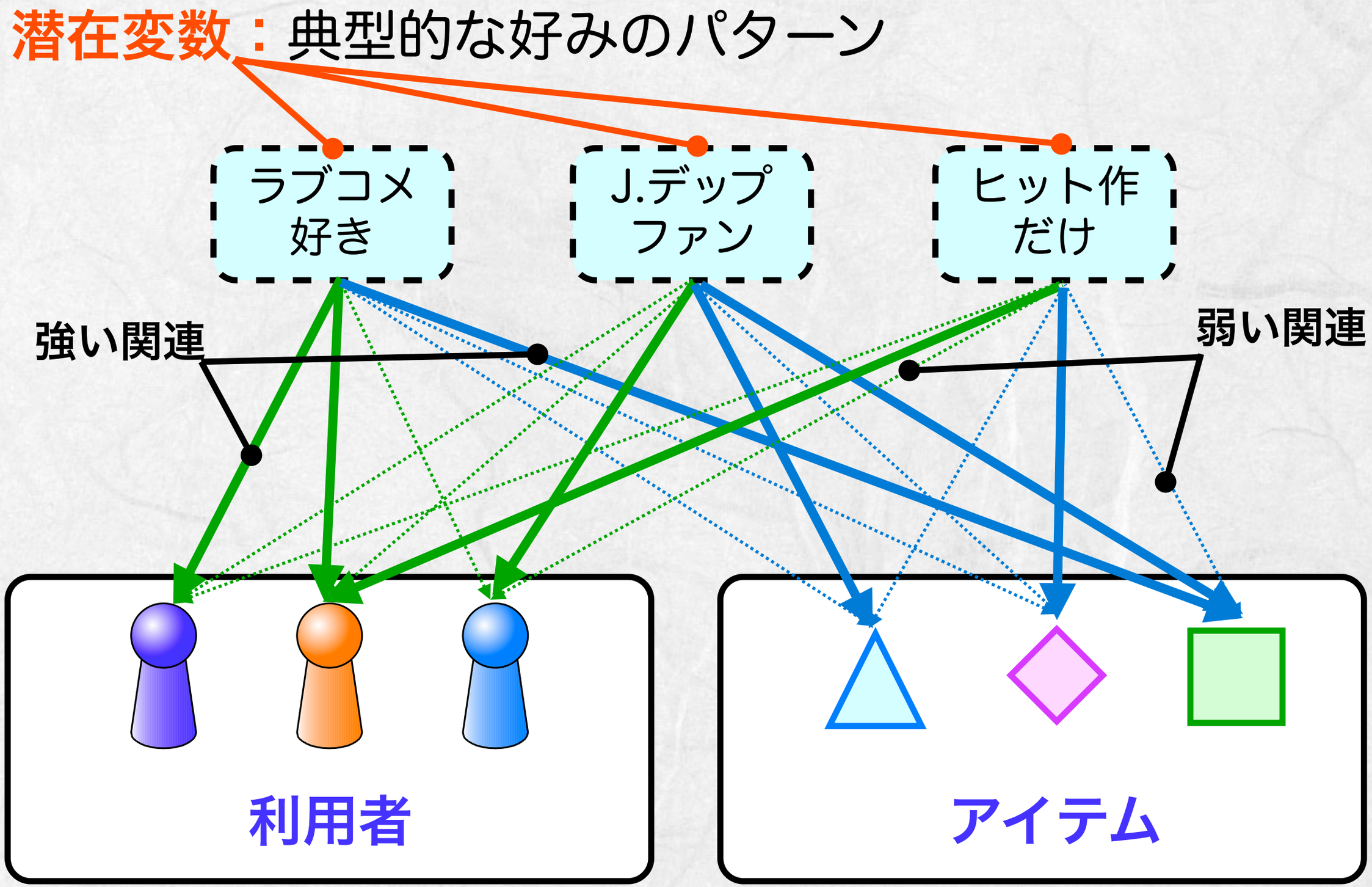
pLSA (probabilistic Latent Semantic Analysis)

自然言語処理用の次元削減法 [Hoffmann UAI99] を協調フィルタリングにも適用 [Hoffman+ IJCAI99]．aspectモデルともいう．

- 未評価と否定的評価の区別がなくても適用しやすい
  「購入していない」は知らなくて買っていないのか，嫌いだから買っていないのか分からない
- 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
  アイテム数が多いときには履歴条件型より特に有利
- 並列計算が容易
  pLSAの解法であるEMアルゴリズムの特長

pLSAモデルのイメージ

潜在変数：典型的な好みのパターン

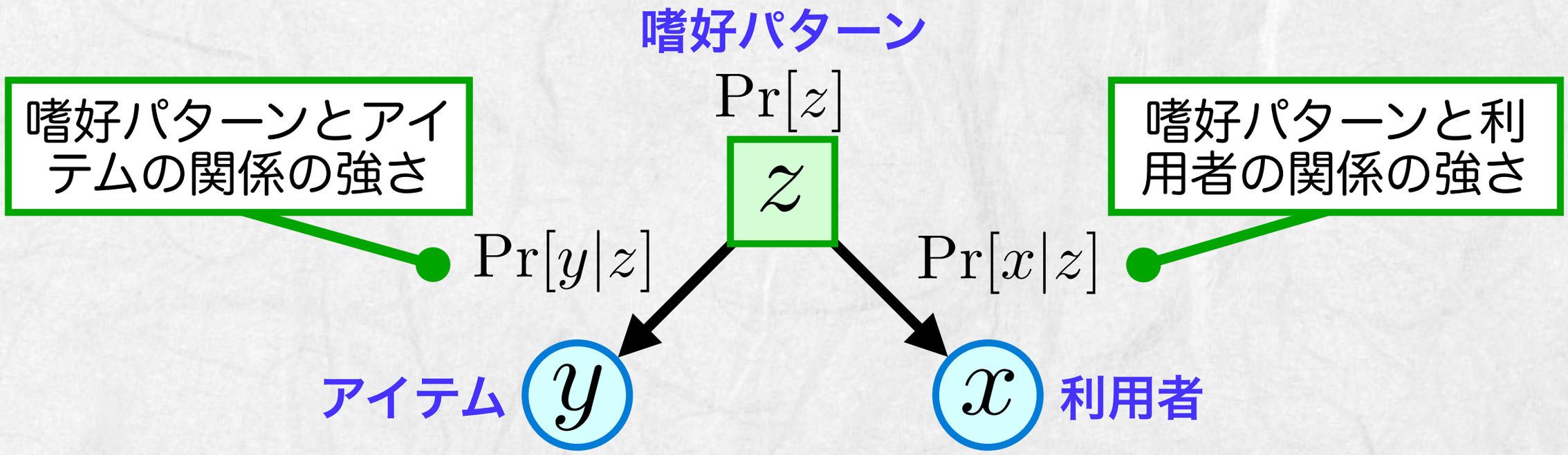


pLSA (1)

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

利用者：　$x \in \{1, \ldots, n\}$
アイテム：$y \in \{1, \ldots, m\}$
潜在変数：$z \in \{1, \ldots, l\}$ (嗜好パターン)



pLSA (2)

pLSAのEMアルゴリズムによる最尤推定
対数尤度関数を最大化して同時確率 $\Pr[x,y]$ を推定

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

Eステップ 現在のパラメータが与えられたときの潜在変数の事後確率

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

Mステップ 現在の潜在変数の分布の下でのパラメータの値

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

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

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

訓練データ中である利用者 x が，アイテム y を購入・閲覧する回数

x は1〜n，y は 1〜m の範囲で計算

尤度関数が収束するまで，EステップとMステップを交互に反復

pLSA

推薦： $x = i$ のときの事後確率を最大化するアイテム

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

事後確率分布関数で $x=i$ を代入

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

学習したパラメータ

事後確率分布： $\Pr[y|x] = \dfrac{\Pr[x, y]}{\sum_y \Pr[x, y]}$

\#
\#

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Mステップの計算

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

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

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

サイト2からサイト1に送信

[Das+ WWW07]

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

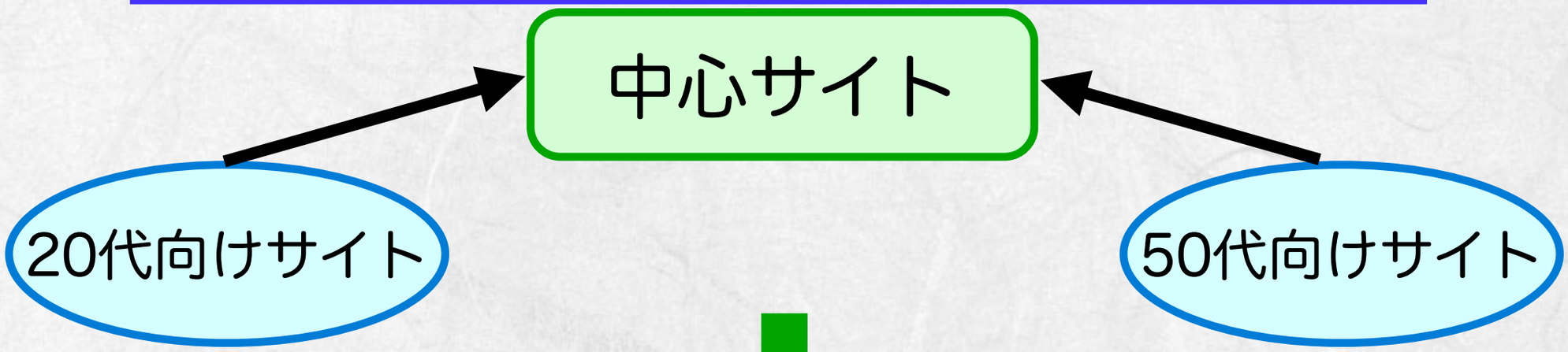
個人情報かどうか?

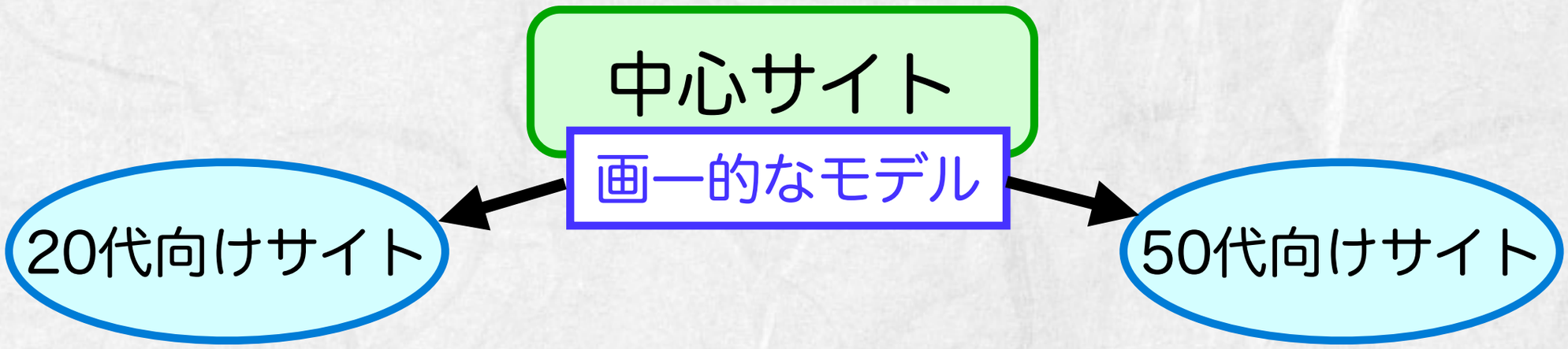
$\Pr[z]$ と $\Pr[y|z]$ は共に x とは無関係なので個人情報ではない

$\Pr[x|z]$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





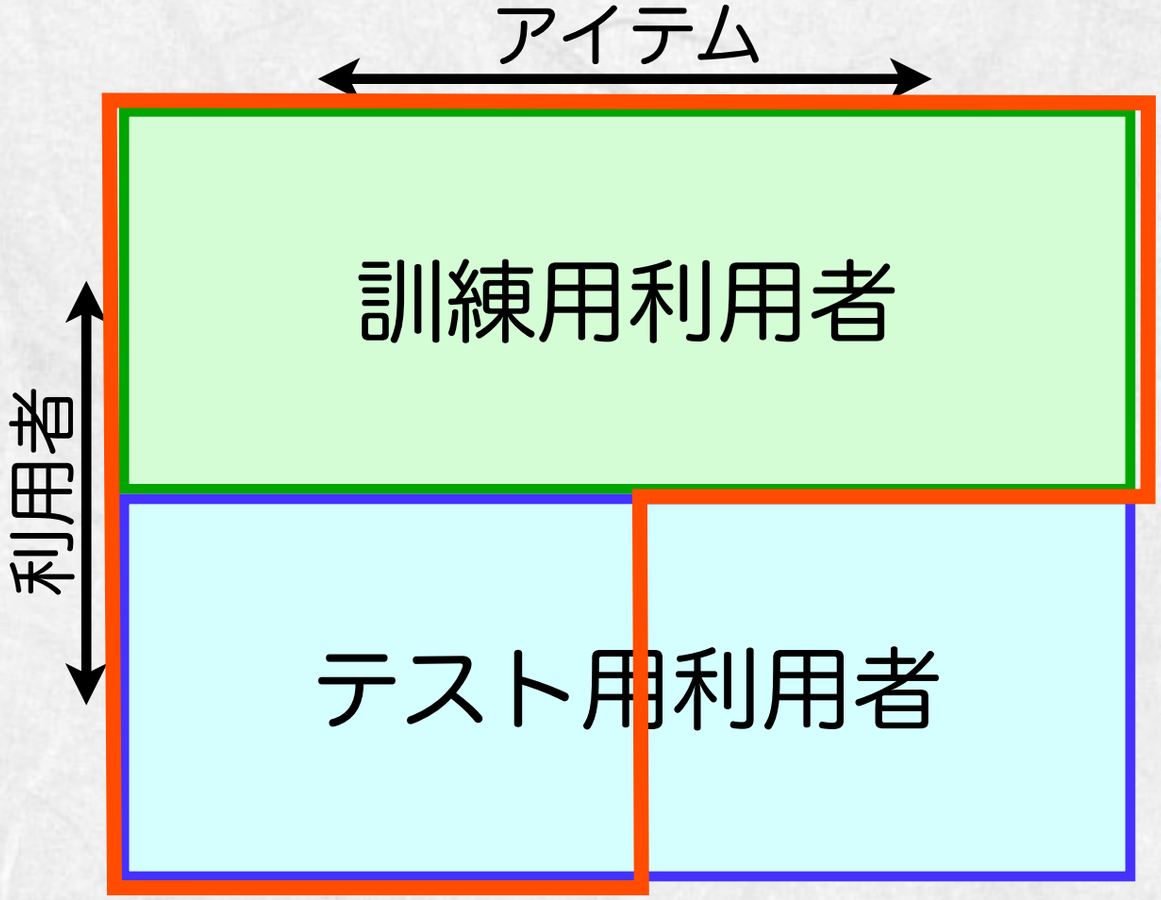
サイトの利用者の偏りが推薦に反映されない

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

他の年代のデータから，各年代の利用者の嗜好を予測

テスト集合	平均確率	人数
全体	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

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

EMアルゴリズムの各反復での十分統計量の計算は，各データについての統計量の和なので容易に並列化可能

↓

- 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

転移学習 (transfer learning)

知識データ

元ドメイン
source domain

- 解きたい問題と似てはいるが，**解きたい問題ではない**
- 解くためのデータや知識は**十分**

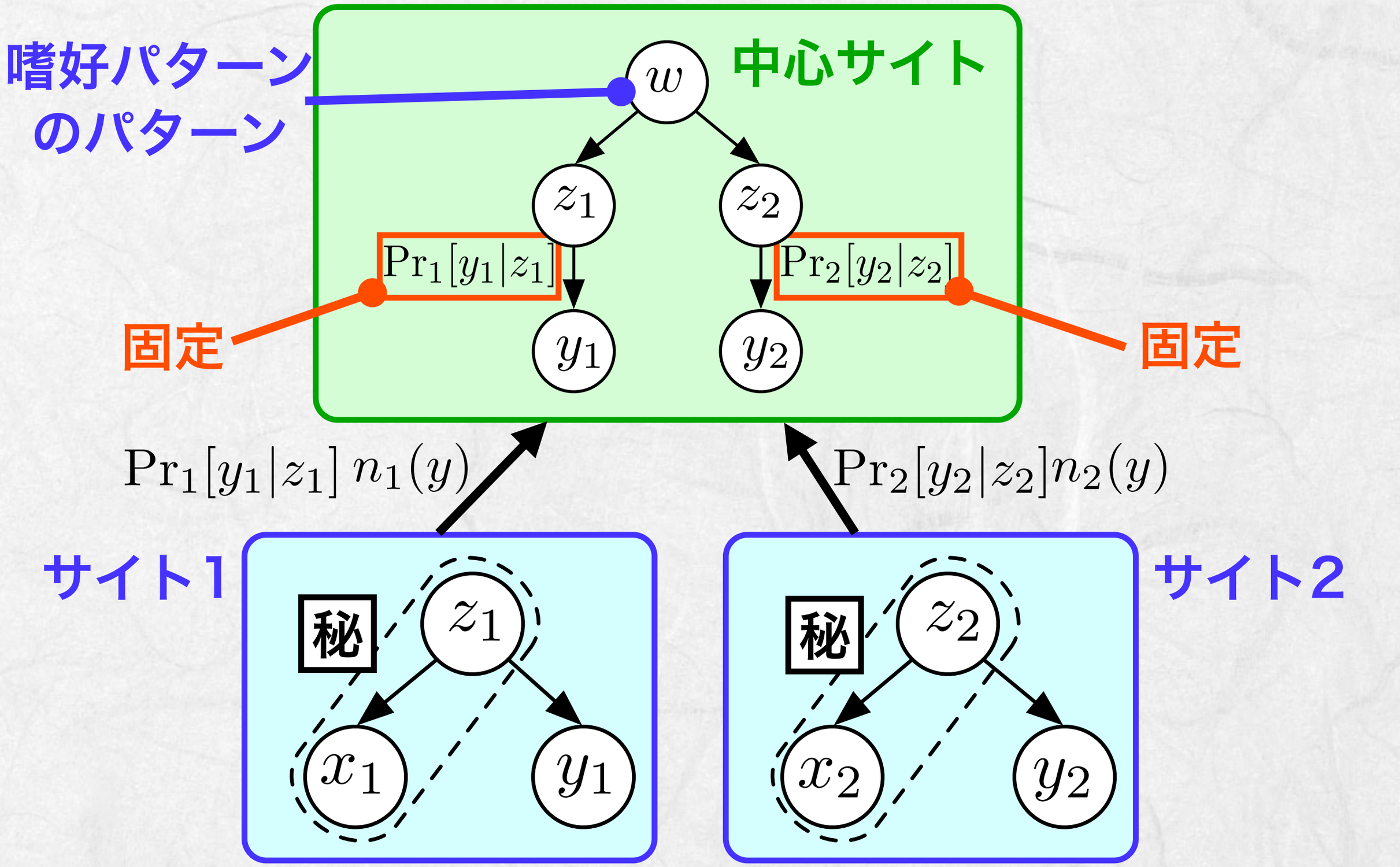
目標ドメイン
target domain

- これから**解きたい問題**
- 解くためのデータや知識は**不十分**

関連したドメインの知識やデータを転移して
目標ドメインの問題をより高精度で解く

マルチタスク学習 目標ドメインと元ドメインの区別がなく，お互いに，別のドメインの知識を導入してそれぞれ精度を上げる

階層ベイズ型のマルチタスク学習



計算方法

サイト1で評価されたアイテム

$$\Pr[y_1] = \sum_{z_1, w} \Pr_1[y_1|z_1] \Pr[z_1|w]$$

潜在変数が分かった場合の対数尤度

$$\log \mathcal{L}_1 = \sum_{y_1} n_1(y_1) \log \left[\sum_{z_1, w} \Pr_1[y_1|z_1] \Pr[z_1|w] \Pr[w] \right]$$

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

パラメータの一部が固定されている経験ベイズ

局所サイトの事前分布を返す

$$\Pr^{new}[z_k] = \sum_{w} \Pr[z_k|w] \Pr[w]$$

実験結果

- 30歳で上下に分け，さらにそれぞれ二つに分割，合計4グループ
- 嗜好パターンのパターン数も変えた

30未満サイトA		30未満サイトB	
元	0.2043	元	0.2147
2	0.2043	2	0.2146
3	0.2042	3	0.2147
5	0.2043	5	0.2147

30以上サイトA		30以上サイトB	
元	0.1874	元	0.1815
2	0.1874	2	0.1814
3	0.1871	3	0.1816
5	0.1874	5	0.1815

- 元：個別に各サイト独立に計算した場合
- 2〜4：嗜好パターンのパターン数 |w| を2〜4に変えて適用

結果が良くない：今後の計画

- Pr[z] だけの改善では，全体に与える影響が小さい
- サイト間で，同じアイテムに関する嗜好パターンの情報が共有されていない．この情報を使って Pr[y | w] を改善する

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**
  - 分散環境での階層モデルの提案 ➡ 結果は良くなかった

今後の予定

- モデルの改良による予測精度の改善