情報の独立性を強化したトピックモデル

神嶌 敏弘[*1]，赤穂 昭太郎[*1]，佐藤 一誠[*2]

[*1] 産業技術総合研究所，[*2] 東京大学

2015年度人工知能学会全国大会（第29回）@ 函館市，2015.6.1

http://www.kamishima.net/

はじめに

[Romei+ 13]

公正配慮型データマイニング

公正性，差別，中立性，独立性などの潜在的な社会的問題について配慮しつつデータマイニングを行う

↓

独立性強化型トピックモデル

- 指定したセンシティブ情報にできるだけ影響されないようなトピックを抽出する
- センシティブ情報に影響されないトピックの獲得や，センシティブ情報に影響されない分類などに利用できる

✤ 差別配慮型DMとも呼ばれているが，ここでは公正配慮型DMと呼ぶ．これは，差別の英語 discrimination が機械学習の文脈では判別の意味になることと，差別への対処以外の問題への適用も可能であるためである．

目次

- **公正配慮型データマイニング**：差別的決定の防止，中立的な情報の提供，関心のない情報の除外
- **公正性の形式的定義：**基本的な表記，トピックモデルにおける公正性
- **独立性強化型トピックモデル：**pLSA，STI-pLSA (strictly topic-independent pLSA)
- **実験と考察：**Reutersデータへの適用，トピックの例，考察
- **まとめ**

公正配慮型データマイニング

差別的決定の防止

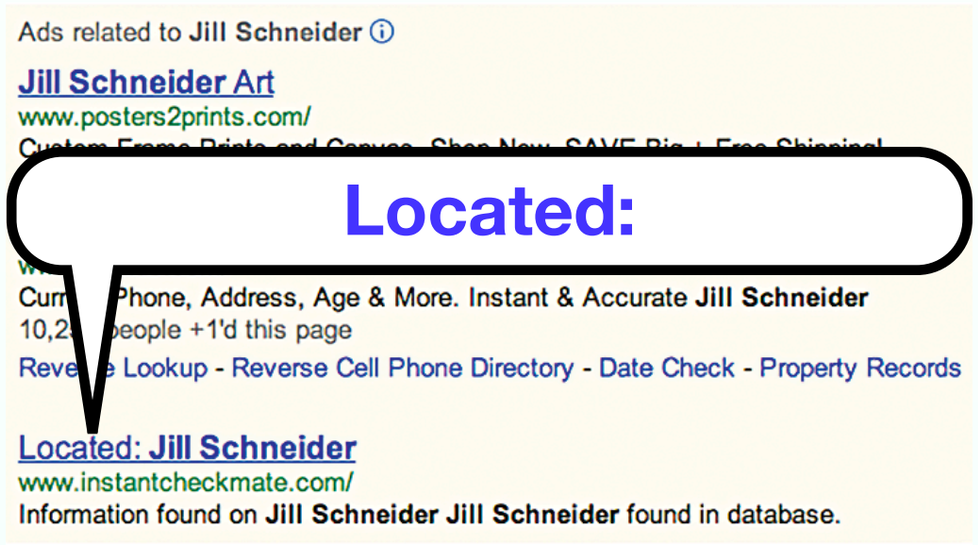
[Sweeney 13]

キーワードマッチ広告配信での懸念

逮捕歴を示唆するような広告文が，ヨーロッパ系で多い名前より，アフリカ系で多い名前でより頻繁に表示された

アフリカ系の名前

ヨーロッパ系の名前



対象者の人種の情報は用いておらず，クリック率向上による副次的な影響によるものであった

このような不公正な決定は公正配慮型DM技術で回避できる

中立的な情報の提供

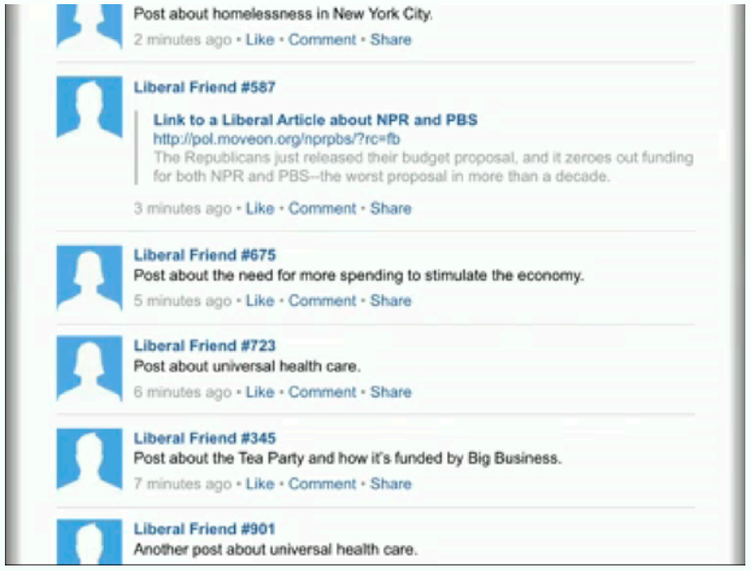
[TED Talk by Eli Pariser, http://www.filterbubble.com/]

フィルターバブル問題

Pariserは，個人化技術により，人々がふれる情報の話題に偏りが生じ，また狭まるとの懸念を示した．

Facebookの友人推薦の例

Pariserの嗜好に合わせて，友人推薦リストから保守派の人が，
知らされない間に除外されていた



FADM技術は中立的な情報を提供するのに役立つ

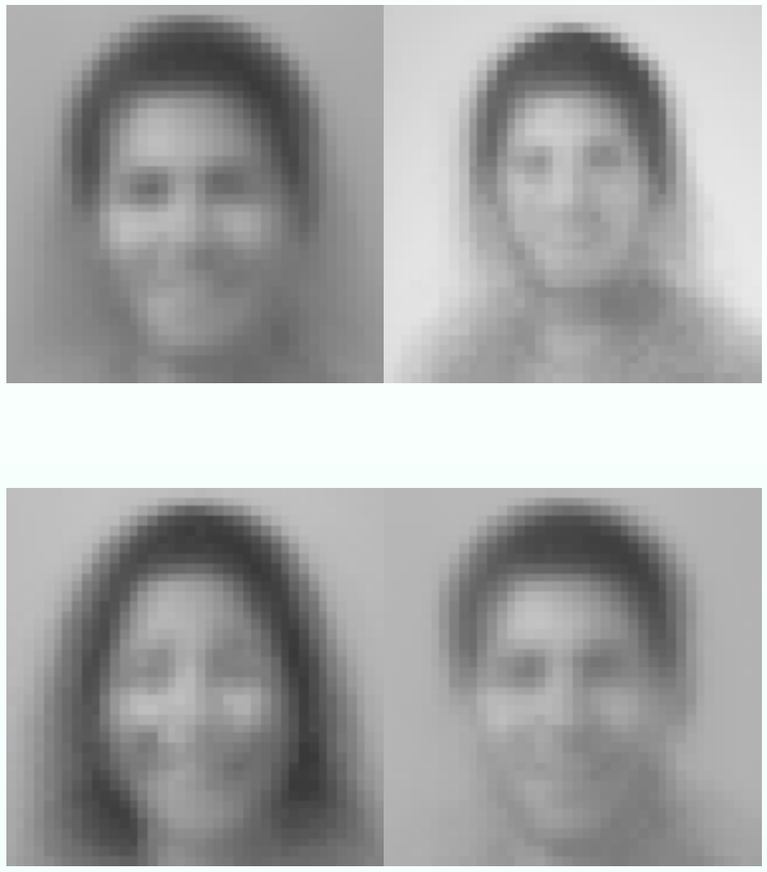
関心のない情報の除外

[Gondek+ 04]

非冗長クラスタリング (non-redundant clustering)

無関心な分割とはできるだけ独立な分割を抽出するクラスタリング

情報ボトルネック法を拡張した，条件付き情報ボトルネック法



顔画像集合のクラスタリング

- 単純にクラスタリングすると，顔だけと，肩も含めた画像に分割された
- 分析者は，こうした分割には意味的に興味深くないと考えた
- この分割とは独立となるようにクラスタリングすると男女のクラスタが得られた

FADM技術により，関心のない不要な情報を除外できる

公正性の形式的定義

基本的な表記

Z トピック変数
objective variable

- これから抽出する文書の話題
- 観測されない潜在変数
- $1 \cdots K$ の多値変数

S センシティブ属性
sensitive attribute

- 影響を除外したい情報
- 2値変数

X 文書変数
document variable

- 文書：単語の集合
- $1 \cdots n$ の多値変数

Y 単語変数
word variable

- 単語：文書を構成する要素
- $1 \cdots m$ の多値変数

トピックモデルにおける公正性

データマイニングにおける公正性

センシティブな情報が決定に影響しない

↓

センシティブ特徴と相関がある非センシティブ特徴は
センシティブ情報を含んでしまっている

↓

red-lining 効果：たとえセンシティブ特徴を利用せずに計算しても，
公正な決定はできない

Z と S は，X や Y が与えられたときに条件付き独立：$Z \perp\!\!\!\perp S \mid X, Y$

↓

センシティブ特徴と目的変数は
無条件に独立である必要　$Z \perp\!\!\!\perp S$

独立性強化型トピックモデル

pLSA

トピック

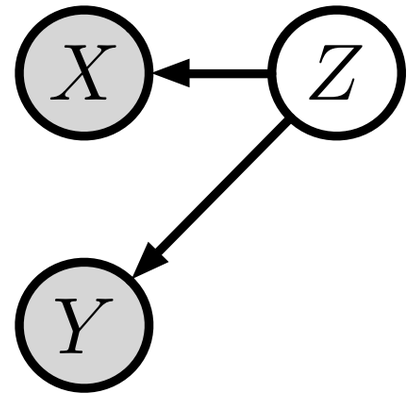
- **気持ち的には：**文章の内容を構成するカテゴリ
- **数理的には：**文章を構成する潜在因子で，その文書中の単語を決定

文書　文書を構成するトピック　トピックに応じて単語生成

$$\Pr[X,Y,Z] = \Pr[X]\ \Pr[X|Z]\ \Pr[Z|Y]$$
$$= \Pr[Z]\ \Pr[Z|X]\ \Pr[Z|Y]$$

計算の都合で変形

グラフィカルモデル表現



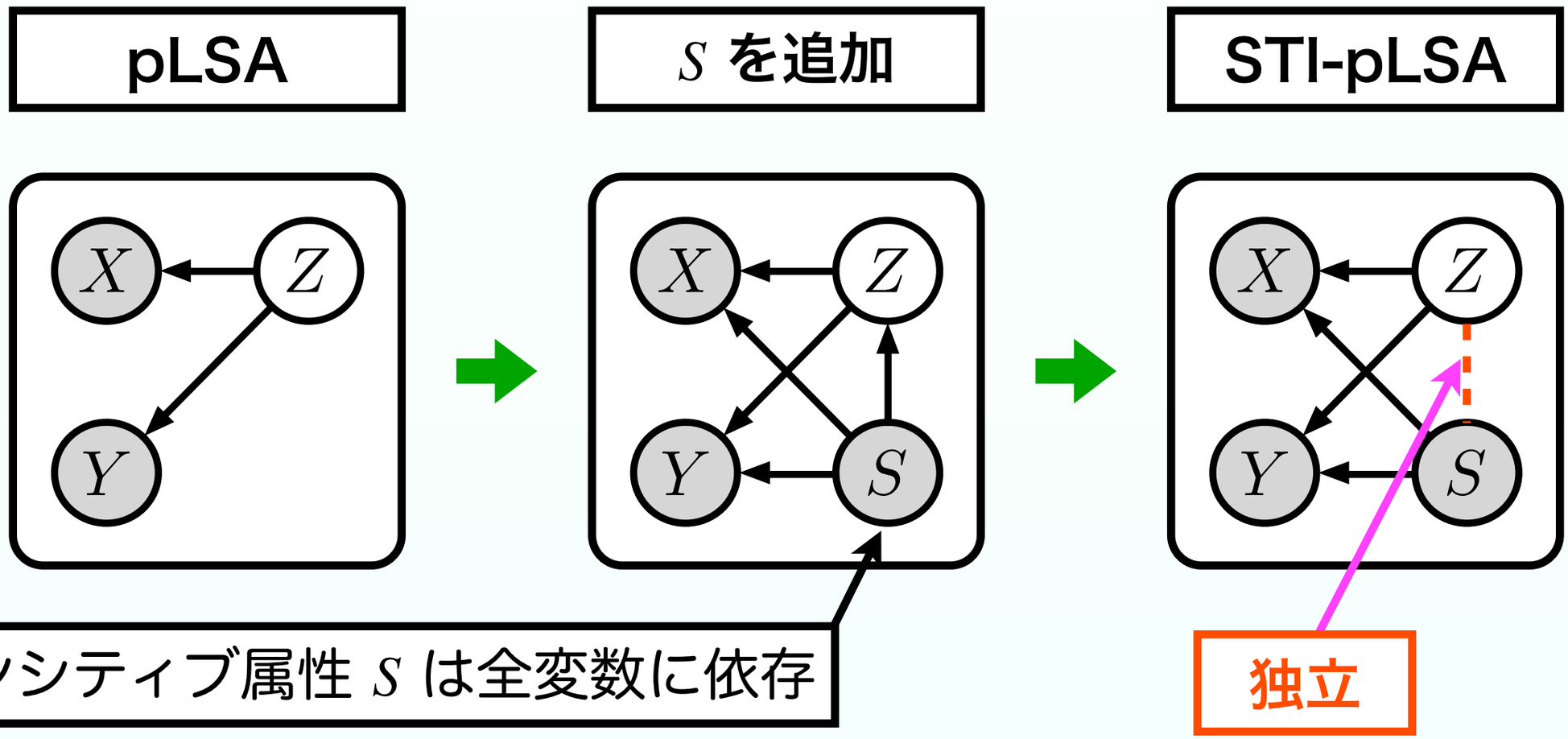
トピックの利用法

- **特徴量：**文 X の内部表現として $\Pr[Z|X]$ を用いる
- **可視化：**各トピックに含まれる語の集合で，文書集合の概要を表示

STI-pLSA

STI-pLSA (strictly topic-independent pLSA)

センシティブ属性 S とトピック Z を独立にした pLSA



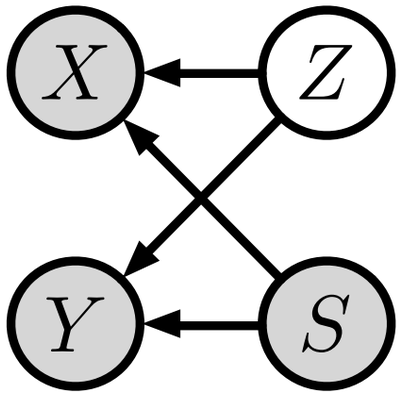
センシティブ属性 S はとトピック Z を独立にする

STI-pLSA

STI-pLSA (strictly topic-independent pLSA)

普通の pLSA と同様にEMアルゴリズムで容易に解ける

全変数の同時分布



$$\Pr[X, Y, Z, S] = \Pr[S]\Pr[Z]\Pr[X|Z, S]\Pr[Y|Z, S]$$

Eステップ

$$\Pr_{\text{new}}[z|x, y, s] \leftarrow \frac{\Pr[z]\Pr[x|z, s]\Pr[y|z, s]}{\sum_{z'}\Pr[z']\Pr[x|z', s]\Pr[y|z', s]}$$

Mステップ

$$\Pr_{\text{new}}[x|z, s] \leftarrow \frac{\sum_{y'} N(x, y', s)\Pr[z|x, y', s] + 1/n}{\sum_{x'y'} N(x', y', s)\Pr[z|x', y', s] + 1}$$

$$\Pr_{\text{new}}[y|z, s] \leftarrow \frac{\sum_{x'} N(x', y, s)\Pr[z|x', y, s] + 1/m}{\sum_{x'y'} N(x', y', s)\Pr[z|x', y', s] + 1}$$

$$\Pr_{\text{new}}[z] \leftarrow \frac{\sum_{x',y',s'} N(x', y', s')\Pr[z|x', y', s'] + 1/K}{N + 1}$$

$$\Pr_{\text{new}}[s] \leftarrow \frac{\sum_{x'y'} N(x', y', s) + 1/2}{N + 1}$$

実験と考察

実験

Reuters-21578 からのトピック抽出

- 出現回数が10回以上の語を抽出 +不要語 や 数字などの削除
- 文書数 $n = 10786$, 単語数 $m = 8210$, トピック数 $K = 2$
- STI-pLSA のセンシティブ属性：pLSA で求めた二つのトピック

pLSA が抽出したトピックの情報に
影響されないトピックを STI-pLSA で抽出する

実験結果

- コーパス中の（文書，語）のペアごとにトピックを予測し，その確率に応じてペアをトピックに割り当てた

S と Z の独立性を χ 二乗検定すると
明らかに独立性は保たれなかった（χ^2=978, 自由度=1）

得られたトピック

pLSA	extinguishment, Diluted, LORD, ABBETT, AUG, Periods, Ended, TELECOMMUNICATIONS, REGULAR, Mths, Rev, Qtly, Oper, Excludes, Shrs, MTHS, Payable, QTLY, mths, shrs, Revs, oper, Qtr, Shr, Avg,	Maekawa, Kahan, Mulford, poorest, Siew, Este, Denman, Papandreou, Yulo, Rafsanjani, Wattari, PHLX, Iranians, Republicans, DIW, Lim, SUMITA, Bennett, ASEAN, vowed, NATO, Clercq, Kiichi, Aegean,
STI-pLSA	Avia, DataCard, Instinet, Sante, Outboard, Resorts, unquoted, Calmat, Metex, Comdata, Cyacq, Gabelli, Cardis, Partly, REPH, Scan, Becor, Sunter, Zico, Symbion, BCW, Quest, Seton, Deak, Gluck, ASSISTANCE, Calny,	LORD, ABBETT, ECUS, BBL, RAINFALL, SAO, BARREL, LTV, POSTINGS, postings, Spectra, Physics, Citgo, Mths, BBLS, Trailways, Oper, AUG, Mthly, extinguishment, Diluted, Ended, Excludes, REBATE, Periods, Evergo,

考察

[Kamishima+ 13]

モデルは独立なのに，実験結果が独立ではないのはなぜか？

モデルから導出される同時分布

$$\Pr[X, Y, Z, S] = \Pr[S] \Pr[Z] \Pr[X|Z, S] \Pr[Y|Z, S]$$

データを分類したときの同時分布

$$\Pr[X, Y, Z, S] = \Pr[Z|X, Y, S] \Pr^*[X, Y, S]$$
$$= \frac{1}{|\mathcal{D}|} \sum_{\mathcal{D}} \Pr[Z|X, Y, S]$$

モデルの分布　　標本分布で近似　　真の分布

モデルバイアス：真の分布とモデルによる分布の乖離 ➡ 無視できない

まとめ

まとめ

- 公正配慮型データマイニングの新たな問題としてトピックモデルを対象とした問題を提案
- STI-pLSAモデル (strictly topic-independent pLSA) を提案し，実験によりその振る舞いの予備的な解析

今後の予定

- モデル上の分布だけでなく，真の分布とのモデルバイアスを考慮した分布との独立性を保つ工夫
- 公正配慮型分類や推薦などの他の応用問題での性能検証

謝辞

- 本研究はJSPS科研費 16700157，21500154，および24500194 の助成を受けたものである