6. 隠れた値を推定する

トピックモデル，クリックモデル，…

梅本 和俊

情報通信研究機構／東京大学
umemoto@tkl.iis.u-tokyo.ac.jp

前回｜主成分分析（とt-SNEとUMAP）

高次元データ is everywhere

10

文書データ（次元: 単語）

画像出典: https://en.wikipedia.org/wiki/Main_Page

画像データ（次元: 画像特徴量）

画像出典: https://www.silhouette-illust.com/illust/15878

医療データ（次元: 疾患, 医薬品, …）

画像出典: https://www.irasutoya.com/2012/12/blog-post_4261.html

ゲノムデータ（次元: 遺伝情報）

画像出典: https://www.irasutoya.com/2015/04/dna.html

次元削減

12

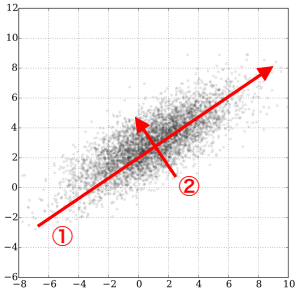
- 入力: 膨大な数 D の次元からなるデータ $x = [x_1, \cdots, x_D]$
- 出力: x を少数 D' $(D' < D)$ の次元で表したデータ $y = [y_1, \cdots, y_{D'}]$

- **成績理解**
  - 入力: [国語の点数, 数学の点数, ...]
  - 出力: [総合成績, 文系／理系]
- **健康評価**
  - 入力: [身長, 体重, ...]
  - 出力: [大きさ, 肥満度, ...]

どうすれば？

14

1. 5つの次元 (車, car, バイク, 図書館, 本) を無理やり2次元にするならどれをまとめる？
   - (車・car・バイク, 図書館・本)
2. 右の2次元データを1次元に削減するならどこに軸を取る？
3. もう1つ軸を加えるならどこに取る？

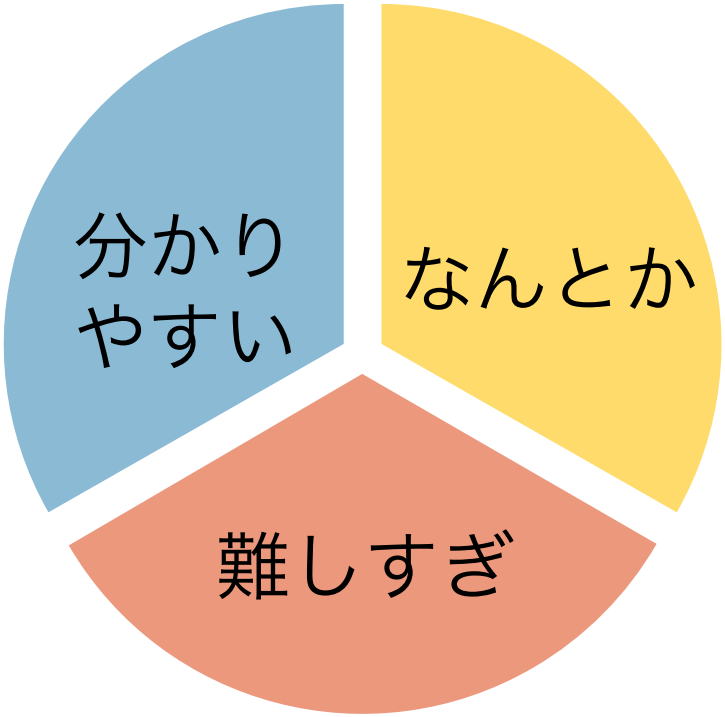


主成分分析｜ポイント

15

- **新次元（主成分）を旧次元の線形結合で作る**
  - $y_{d'} = w_1 x_1 + \cdots + w_D x_D$
  - 乗り物 $= 0.6\times$車 $+ 0.4\times$car $+ 0.3\times$バイク $+ \cdots$
- **主成分間には相関がない**（≒全く別の事柄を表現）
  - OK:「乗り物」成分,「勉強」成分
  - NG:「乗り物」成分,「飛行機」成分
- **データの散らばりが大きくなるように主成分を選択**
  - 分散を最大化→情報量を最大化→情報損失を最小化
- **主成分の寄与率が分かる**
  - この新次元だけで全データの何%を説明可能か

前回のアンケート｜内訳



個人的な目標

- **座学:** 直感的説明を多め＋発展的話題を少し＋脱線をたまに（？）
- **演習:** 基本的なところを確認（今日はちょっと少ない…）

前回のアンケート｜主成分分析

- **説明が速くてP17の数式が追えなかった**
  - 簡単におさらい＆補足

主成分分析｜計算方法｜分散の最大化

17

- **単位ベクトルのノルム制約を考慮した次式を最大化**

$$J = e^{\mathsf{T}} \boldsymbol{V} \boldsymbol{e} - \lambda \left(\boldsymbol{e}^{\mathsf{T}} \boldsymbol{e} - 1 \right)$$

cf. ラグランジュの未定乗数法

- **J を e で偏微分した値が最大時には0となるべきなので**

$$\frac{\partial J}{\partial e} = 2\boldsymbol{V}e - 2\lambda e = 2\left(\boldsymbol{V} - \lambda \boldsymbol{I}\right) e = \boldsymbol{0}$$

固有値問題
- λ は $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}$ の固有値の1つ
- $\boldsymbol{e}$ はそれに対応する固有ベクトル

- **分散に代入すると**

$$e^{\mathsf{T}} \boldsymbol{V} \boldsymbol{e} = \boldsymbol{e}^{\mathsf{T}} (\lambda \boldsymbol{e}) = \lambda$$

第1軸の単位ベクトル $\boldsymbol{e}$

分散共分散行列の**最大固有値** $\boldsymbol{\lambda_1}$ に対応する**固有ベクトル**

前回のアンケート｜主成分分析

- **特徴量が正規分布に従うという仮定？**
  - それ以外でも計算可能だが，うまく分解できないことも [Shlens, 2014]

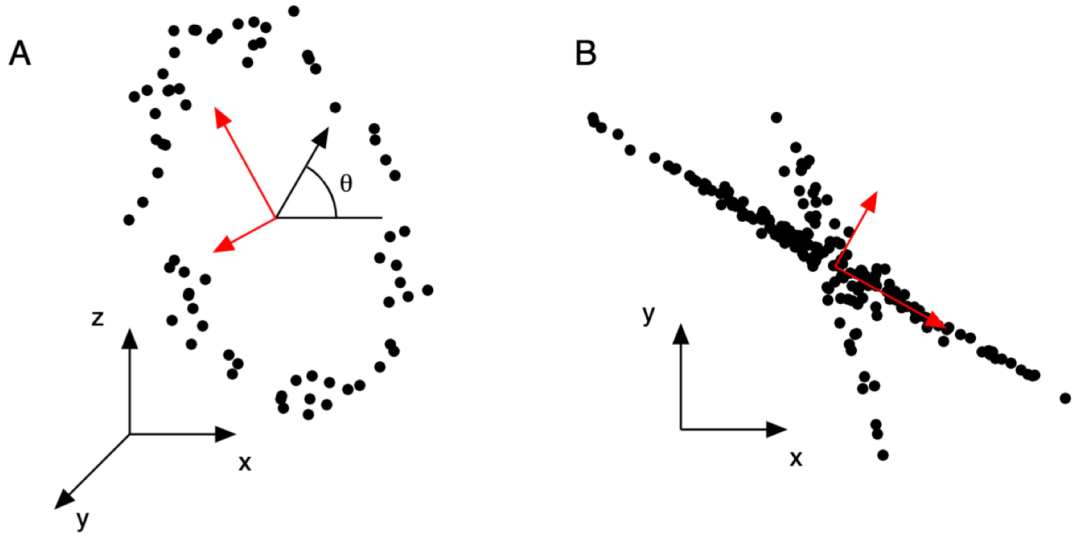


FIG. 6 Example of when PCA fails (red lines). (a) Tracking a person on a ferris wheel (black dots). All dynamics can be described by the phase of the wheel θ, a non-linear combination of the naive basis. (b) In this example data set, non-Gaussian distributed data and non-orthogonal axes causes PCA to fail. The axes with the largest variance do not correspond to the appropriate answer.

前回のアンケート｜こんな話を聞きたい（1）

- **情報検索（IR）に関する話題**
  - ところどころに混ぜます
- **（統計的？）因果推論**
  - 私も勉強中です…
  - https://www.amazon.co.jp/dp/4254122411
  - IR: 位置バイアスを含むクリックログからの反実仮想ランキング学習
- **現場で起きた話**
  - （現場じゃないけど）医療解析の話を少し紹介します

前回のアンケート｜こんな話を聞きたい（2）

- **勾配ブースティング**
  - 詳しく知りません…
  - https://www.slideshare.net/itakigawa/ss-77062106
  - 1個の決定木→複数の木のアンサンブル（独立: バギング, 順番: ブースティング）
  - IR: LambdaMARTという（確か）勾配ブースティングに基づくランキング学習が一時流行

- **Kaggleで使える実践的なテクニック**
  - やったことない…
  - 勾配ブースティング系がよく使われるイメージ（チューニング・解釈のしやすさ）

前回のアンケート｜データサイエンス

- **お薦めサイト・参考書**
  - あまり読んでいないので分からない…
- **便利なツール**
  - 言語: 書き（読み）やすさ, インタラクティブ性, ライブラリの豊富さを総合するとPython？
  - 定番ライブラリ: jupyter, ipython, scikit-learn, numpy, scipy, pandas, ...
  - あとはタスクに応じて使い分け
- **これをすべし**
  - 実データに触れる
  - モデルを学習する以前にデータの性質を理解することが超大事
  - 国際会議SIGMOD2019のキーノートを少し紹介します

Lise Getoor
(UC Santa Cruz)

Keynote 1

Responsible Data Science

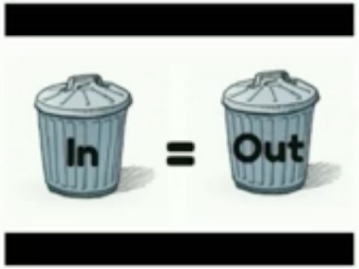
Abstract. Data science is an emerging discipline that offers both promise and peril. Responsible data science refers to efforts that address both the technical and societal issues in emerging data-driven technologies. How can machine learning and database systems reason effectively about complex dependencies and uncertainty? Furthermore, how do we understand the ethical and societal issues involved in data-driven decision-making? There is a pressing need to integrate algorithmic and statistical principles, social science theories, and basic humanist concepts so that we can think critically and constructively about the socio-technical systems we are building. In this talk, I will overview this emerging area, with an emphasis on relational learning.

https://sigmod2019.org/sigmod_keynote

データに潜むバイアス

DATA BIAS

If the input to the system is biased,
due to selection bias, institutional bias, or societal bias,
then the output will be biased

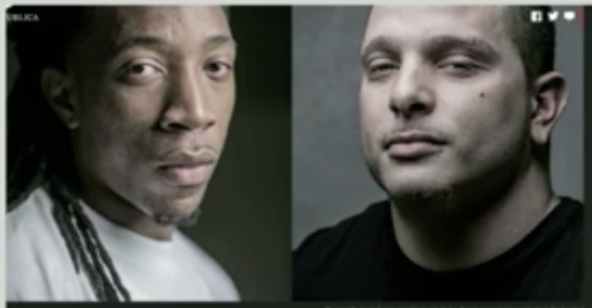


Famous Computer Science phrase: GARBAGE IN, GARBAGE OUT

再犯予測

RECIDIVISM RISK defendant's likelihood of committing a crime

USED THROUGHOUT CRIMINAL JUSTICE SYSTEM pretrial, bail and sentencing



Machine Bias

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes.
(Source: ProPublica analysis of data from Broward County, Fla.)

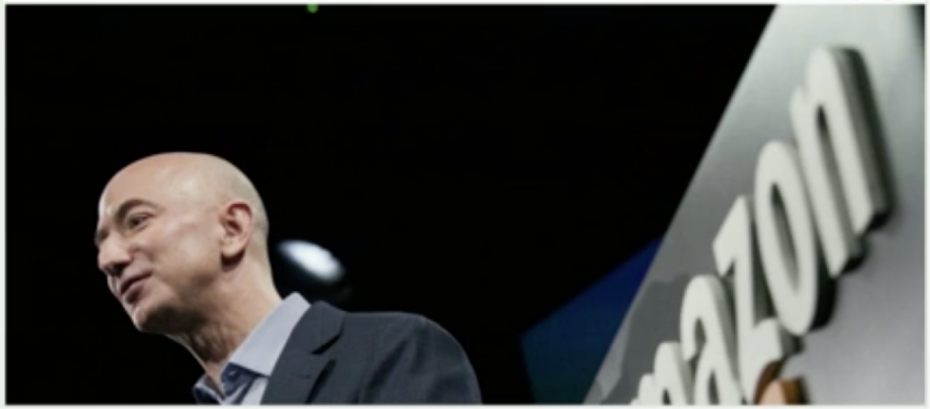
Overpredicts recidivism for African Americans; underpredicts recidivism for whites

underpr Propublica, May 2016

職員採用

Amazon built an AI tool to hire people but had to shut it down because it was discriminating against women

Isobel Asher Hamilton Oct. 10, 2018, 5:47 AM



Between 30 and 50 percent of programmers were women in the 1950s, and it was seen as natural career for them, as evidenced by a 1967 *Cosmopolitan* feature about "Computer Girls."

"It's just like planning a dinner…You have to plan ahead and schedule everything so that it's ready when you need it," Dr. Hopper told the magazine. "Women are 'naturals' at computer programming."

TIMELINE - Medium.com

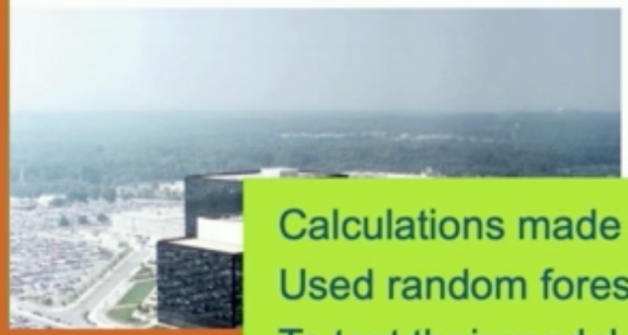
テロリスト認識

MILITARY

Can The NSA's Machines Recognize A Terrorist?

The big problem with little data

By Dave Gershgorn February 16, 2016



Calculations made on 80 variables for each cell phone user

Used random forest

To test their model, project uses data from just 7 known terrorists, plus random sample of 100,000 mobile phone users

When run it the wild, it identified a Al Jazeera reporter covering Al Qaeda as a potential terrorist

狼 or ハスキー犬？

DEEP LEARNING

Deep learning, or **deep neural networks,** build a abstract, simplified neural network.

Given a bunch of observed input/output pairs, they are really good at constructing an abstract representation.

The models are black-box – it is difficult to interpret what the models are capturing.

(a) Husky classified as wolf (b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

Table 2: "Husky vs Wolf" experiment results.

キーノート（の導入部分）の紹介はここまで

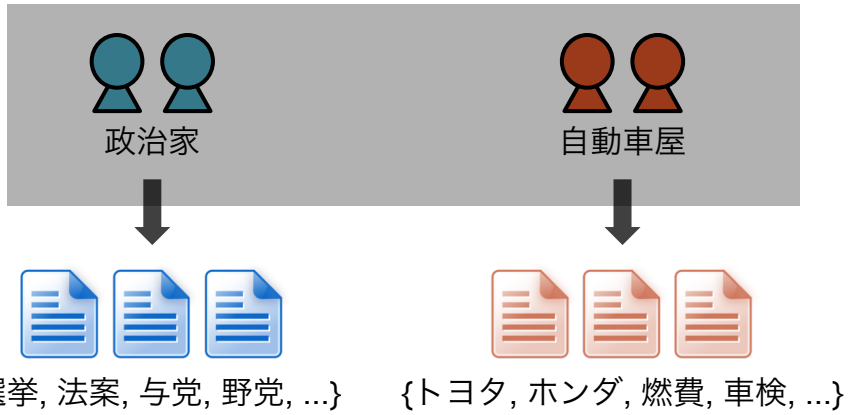
- **言いたかったこと（想像）**
  - 高い精度さえ出せば何でも良いってわけではない
    - なぜ上手くいく／いかない？
    - 社会的／倫理的な影響は？
  - （責任あるデータサイエンスのためにDB研究者は何ができる？）

- **講演動画がここで見られます**
  - https://av.tib.eu/media/42856

- **NECの小山田さんによるSIGMOD2019の素晴らしいまとめ**
  - https://www.slideshare.net/stillpedant/sigmod-2019
    - P28 – P61がこのキーノートの部分
    - その他の部分にもデータに関する研究の最前線が凝縮されている

閑話休題｜観測データの背後に潜む隠れた情報

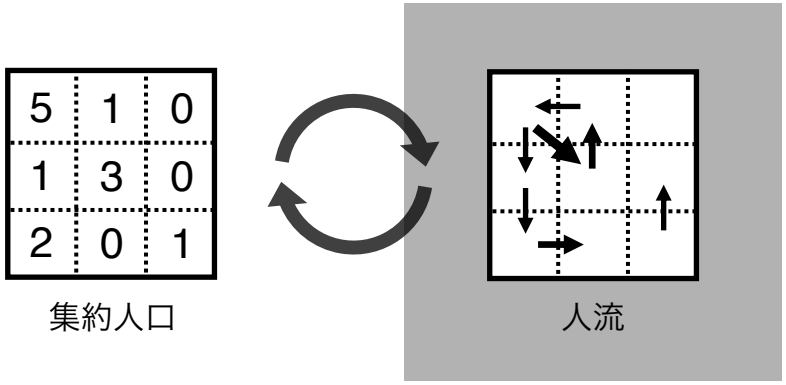
同じトピックの文書では似た単語が出現しやすい



共通した嗜好を持つ人は同じ商品を買いやすい

検索結果ページでのクリック行動の要因

人の移動によって各地の混雑度が時間変化する



潜在変数モデル

隠れた情報（潜在変数 z）からデータ x が生成される過程を
条件付き確率 $p(x|z)$ でモデル化

基本的な流れ

1. モデル

潜在変数やデータの生成過程の記述に適切な確率分布を決定

2. 学習

潜在変数と確率分布パラメータを観測データから教師なしで推定

3. 応用

クラスタリング，検索，推薦，…（未知データにも対応可能なモデルも）

潜在変数モデル｜ステップ1（モデル）

潜在変数やデータの生成過程の記述に適切な確率分布を決定

潜在変数（例: トピック） z

$z = 1$ 政治

$z = 2$ 自動車

$z = \cdots$ …

確率分布 $p(z)$

トピック z が
どれくらい発生しやいか

データ（例: 文書） x

確率分布 $p(x|z)$

トピック z で文書 x が
どれくらい発生しやすいか

データ x の生成確率 $p(x) = \sum_z p(z)p(x|z)$

潜在変数モデル｜ステップ3（応用）

ステップ2（後述）により $\{p(z)\}_z$ と $\{p(x|z)\}_{z,x}$ が推定できた

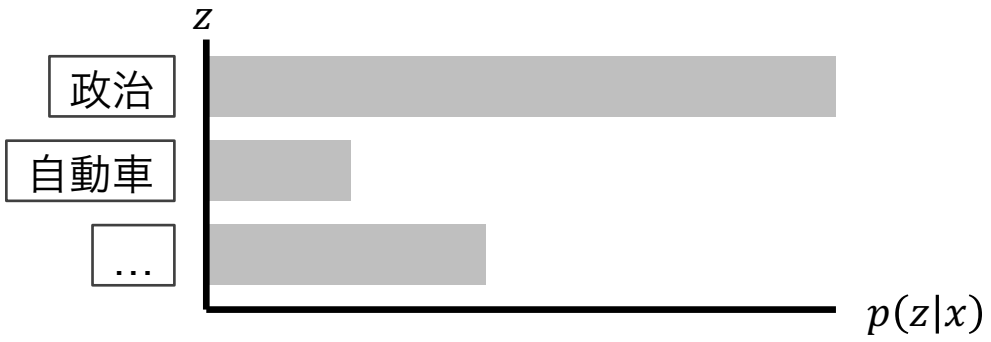
ベイズの定理

$$p(z|x) = \frac{p(z,x)}{p(x)} = \frac{p(z)p(x|z)}{\sum_{z'} p(z',x)} = \frac{p(z)p(x|z)}{\sum_{z'} p(z')p(x|z')}$$

（既知 or 未知）データ x の潜在変数の分布は…

文書 x

{選挙, 法案, 与党, 野党, ...}



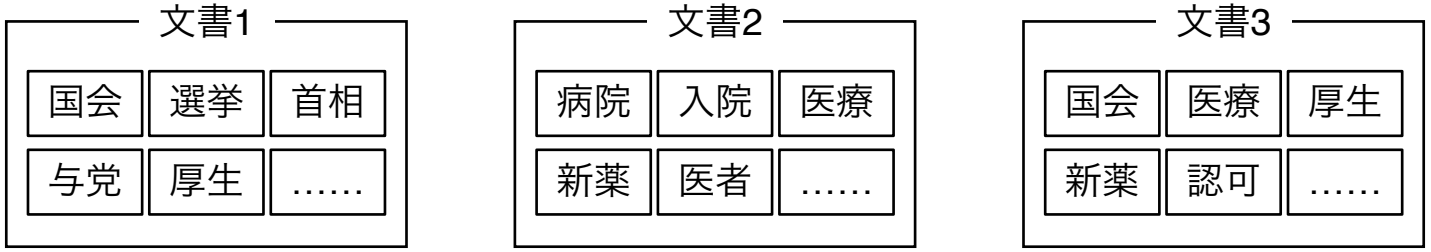
応用例

- 同じ話題に関する文書集合をまとめる
- キーワードと（表層的には異なるが）意味的にマッチする文書を検索する
- 似た嗜好を持つユーザが好む商品を推薦する

トピックモデル

トピックモデル

文書（単語の集合）の生成過程を記述する統計モデル



単純

ユニグラムモデル

全ての文書が同一のトピックに所属することを仮定した生成モデル

混合ユニグラムモデル

各文書が1つのトピックに所属することを仮定した生成モデル

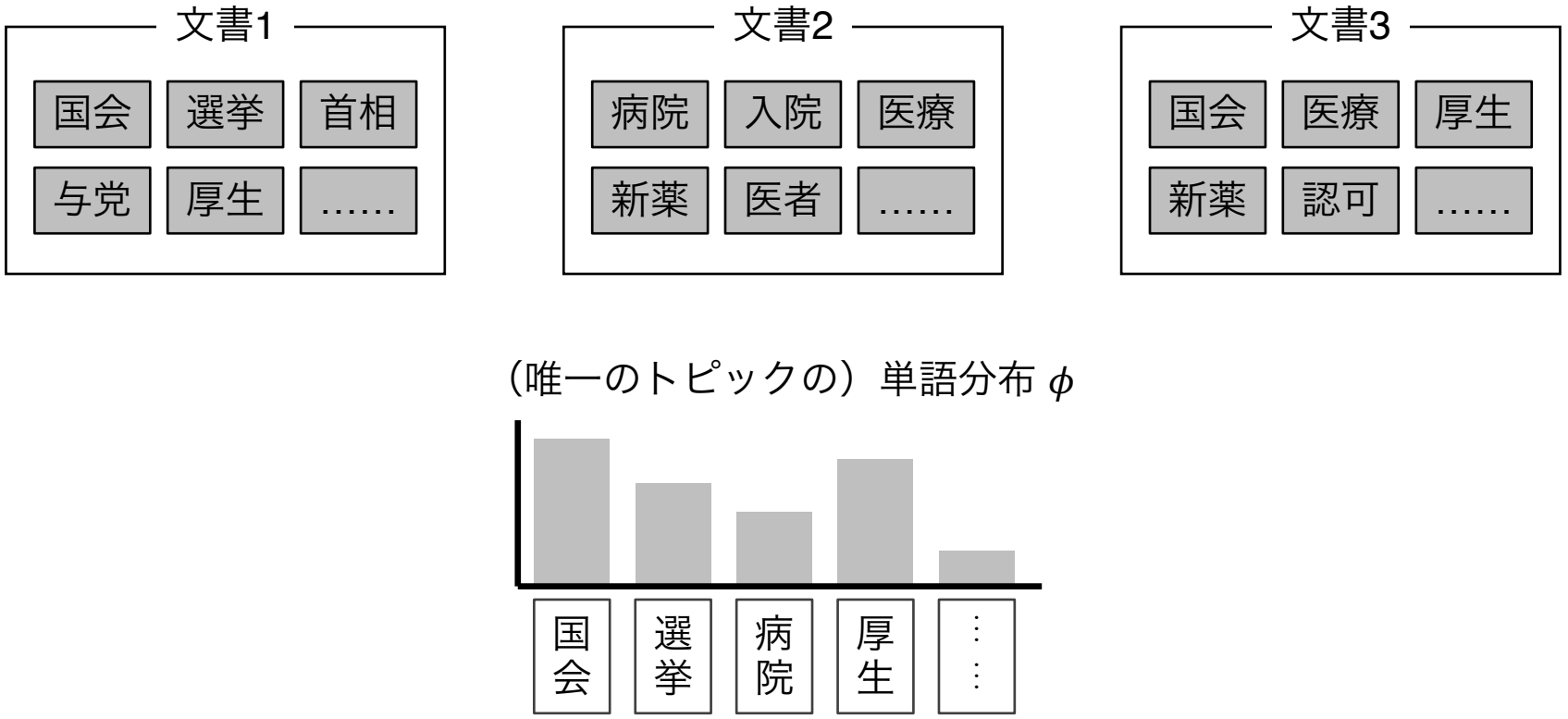
確率的潜在意味解析モデル（PLSA）

複雑

各文書が複数のトピックに所属することを仮定した生成モデル

ユニグラムモデル

全ての文書が同一のトピックに所属することを仮定した生成モデル



ユニグラムモデル｜最尤推定

単語分布に関する制約 $\sum_{v=1}^{V} \phi_v = 1$ の下で

文書集合 W を観測する確率（パラメータの関数と見た場合に尤度と呼ぶ）

$$p(W|\phi) = \prod_{d=1}^{D} p(w_d|\phi) = \prod_{d=1}^{D} \prod_{n=1}^{N_d} p(w_{dn}|\phi) = \prod_{d=1}^{D} \prod_{n=1}^{N_d} \phi_{w_{dn}} = \prod_{v=1}^{V} \phi_v^{N_v}$$

を最大化するパラメータ ϕ^* を発見したい

$$\left\{ \begin{array}{l} f(\phi) = \log p(W|\phi) \\ g(\phi) = \sum_{v=1}^{V} \phi_v - 1 \end{array} \right\}$$ としてラグランジュの未定乗数法を使うと…

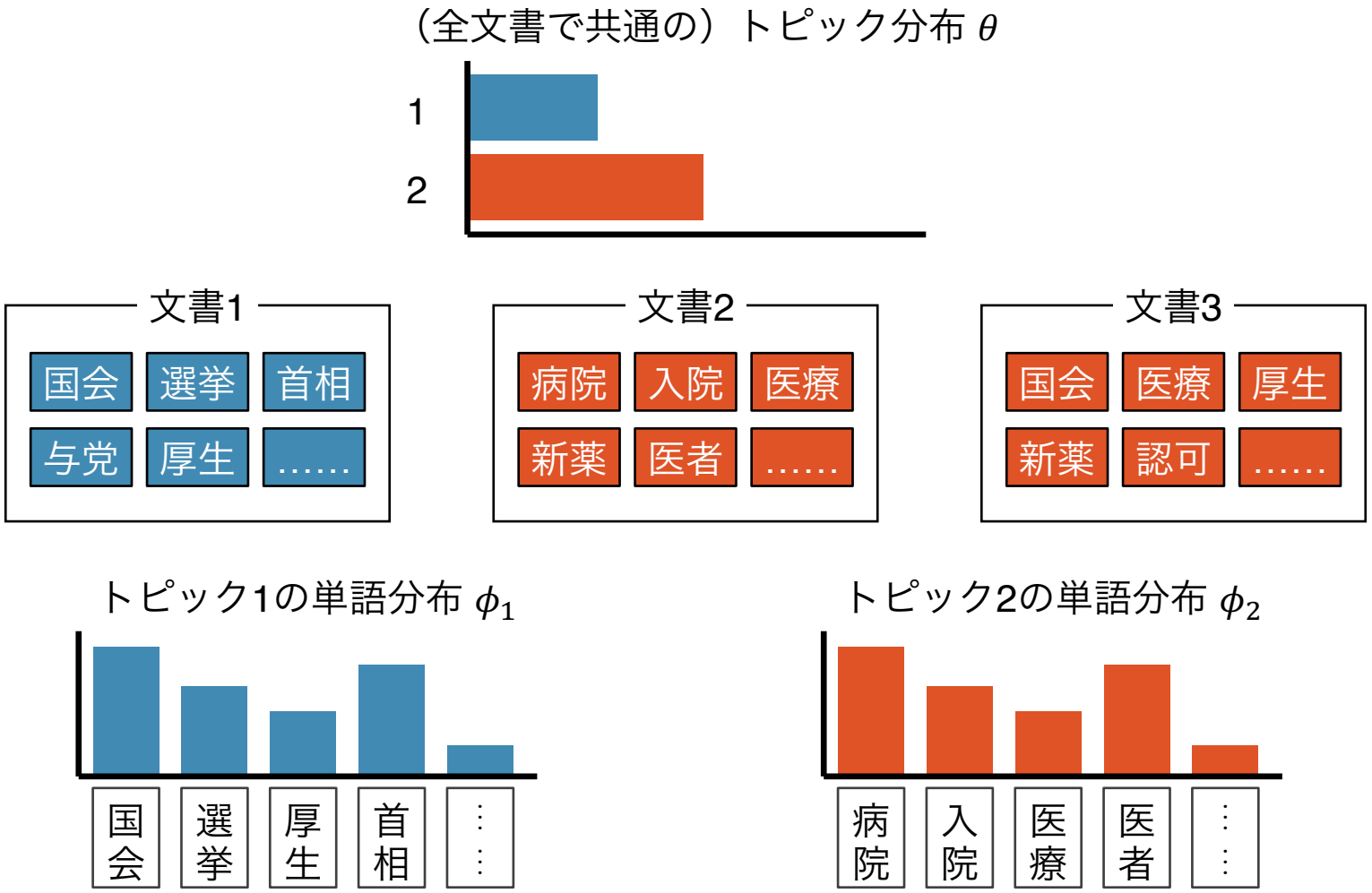
$$\phi^* = \arg\max_{\phi} \mathcal{L} = \arg\max_{\phi} f(\phi) - \lambda g(\phi) = \sum_{v=1}^{V} N_v \log \phi_v - \lambda \left(\sum_{v=1}^{V} \phi_v - 1 \right)$$

さらに $\frac{\partial \mathcal{L}}{\partial \phi_v} = 0$ と $\sum_{v=1}^{V} \phi_v = 1$ を使うと最尤推定量は…

$$\phi_v^* = \frac{N_v}{\sum_{v'=1}^{V} N_{v'}}$$

混合ユニグラムモデル

各文書が1つのトピックに所属することを仮定した生成モデル



混合ユニグラムモデル｜最尤推定

maximize

$$p(W|\theta,\phi) = \prod_{d=1}^{D} p(w_d|\theta,\phi) = \prod_{d=1}^{D}\sum_{k=1}^{K} p(z_d=k, w_d|\theta,\phi)$$

$$= \prod_{d=1}^{D}\sum_{k=1}^{K} p(z_d=k|\theta)p(w_d|\phi_k) = \prod_{d=1}^{D}\sum_{k=1}^{K}\theta_k \prod_{n=1}^{N_d}\phi_{kw_{dn}} = \prod_{d=1}^{D}\sum_{k=1}^{K}\theta_k \prod_{v=1}^{V}\phi_{kv}^{N_{dv}}$$

subject to $\sum_{k=1}^{K}\theta_k = 1$ **and** $\sum_{v=1}^{V}\phi_{kv} = 1 \quad (k=1,\ldots,K)$

- 対数尤度を取るとlogの中にsumが入ってしまい解析解が得られない…
- EMアルゴリズム（省略）で局所最適解を反復的に推定
- q_{dk} は文書 d がトピック k に属する確率（負担率）

Eステップ

$$q_{dk} = \frac{\theta_k \prod_{v=1}^{V}\phi_{kv}^{N_{dv}}}{\sum_{k'=1}^{K}\theta_{k'}\prod_{v=1}^{V}\phi_{k'v}^{N_{dv}}}$$

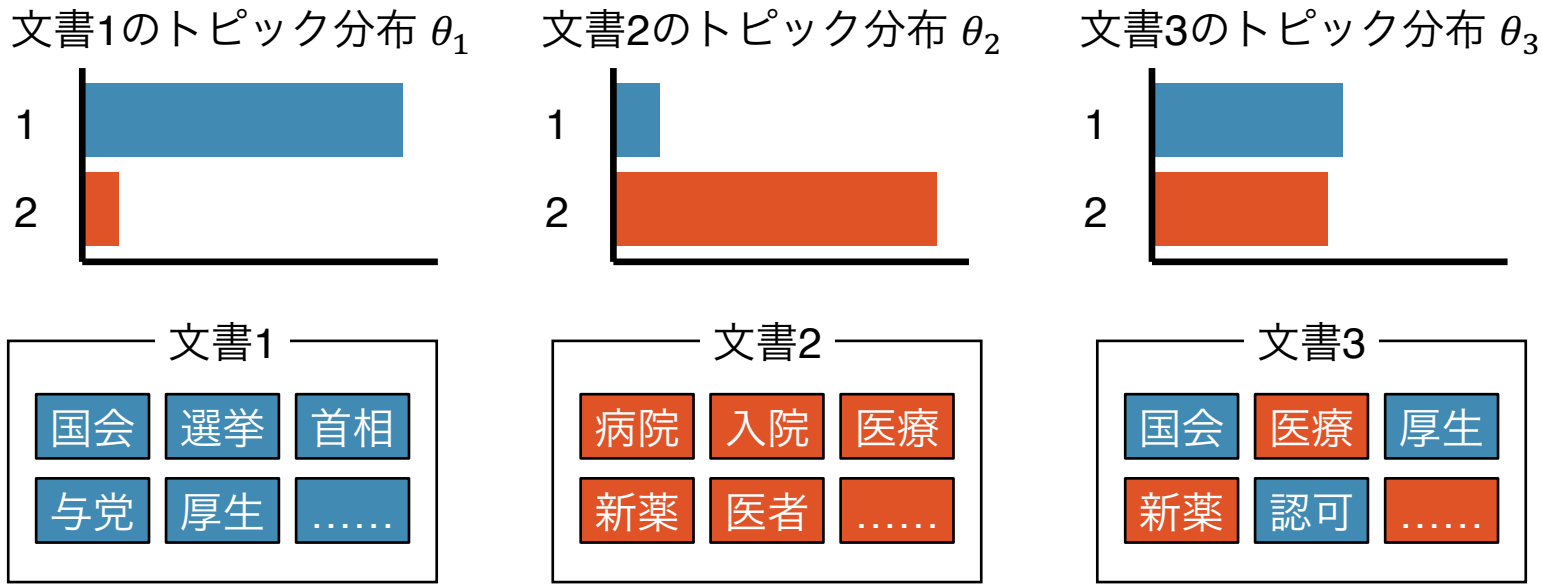
Mステップ

$$\theta_k = \frac{\sum_{d=1}^{D} q_{dk}}{\sum_{k'=1}^{K}\sum_{d=1}^{D} q_{dk'}}$$

$$\phi_{kv} = \frac{\sum_{d=1}^{D} q_{dk}N_{dv}}{\sum_{v'=1}^{V}\sum_{d=1}^{D} q_{dk}N_{dv'}}$$

確率的潜在意味解析モデル（PLSA）

各文書が複数のトピックに所属することを仮定した生成モデル



確率的潜在意味解析モデル（PLSA）｜最尤推定

maximize

$$p(W|\theta,\phi)=\prod_{d=1}^{D}p(w_d|\theta_d,\phi)=\prod_{d=1}^{D}\prod_{n=1}^{N_d}\sum_{k=1}^{K}p(z_{dn}=k,w_{dn}|\theta_d,\phi_k)$$

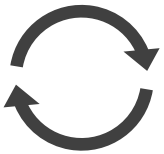
$$=\prod_{d=1}^{D}\prod_{n=1}^{N_d}\sum_{k=1}^{K}p(z_{dn}=k|\theta_d)p(w_{dn}|\phi_k)=\prod_{d=1}^{D}\prod_{n=1}^{N_d}\sum_{k=1}^{K}\theta_{dk}\phi_{kw_{dn}}$$

subject to $\sum_{k=1}^{K}\theta_{dk}=1 \quad (d=1,\ldots,D)$ **and** $\sum_{v=1}^{V}\phi_{kv}=1 \quad (k=1,\ldots,K)$

- 対数尤度を取るとlogの中にsumが入ってしまい解析解が得られない…
- EMアルゴリズム（省略）で局所最適解を反復的に推定
- q_{dnk} は文書 d の n 番目の単語 w_{dn} がトピック k に属する確率（負担率）

Eステップ

$$q_{dnk}=\frac{\theta_{dk}\phi_{kw_{dn}}}{\sum_{k'=1}^{K}\theta_{dk'}\phi_{k'w_{dn}}}$$

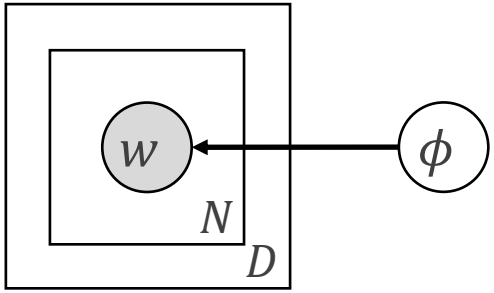


$$\theta_{dk}=\frac{\sum_{n=1}^{N_d}q_{dnk}}{\sum_{k'=1}^{K}\sum_{n=1}^{N_d}q_{dnk'}}$$

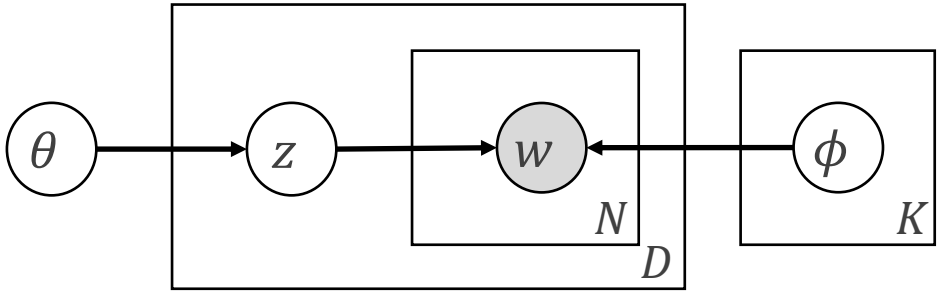
$$\phi_{kv}=\frac{\sum_{d=1}^{D}\sum_{n:w_{dn}=v}q_{dnk}}{\sum_{v'=1}^{V}\sum_{d=1}^{D}\sum_{n:w_{dn}=v'}q_{dnk}}$$

トピックモデル｜グラフィカルモデル表現

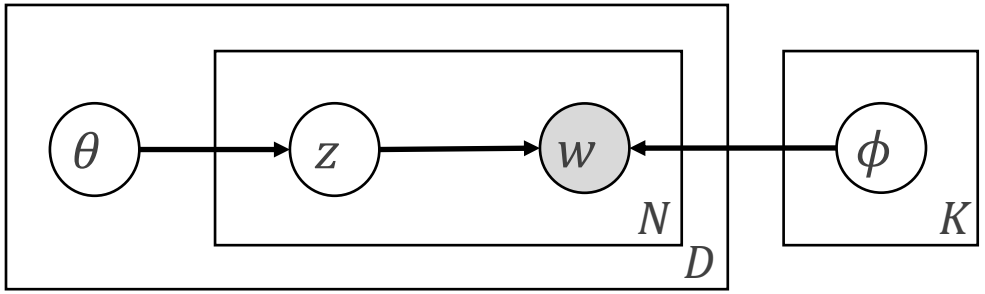
ユニグラムモデル



混合ユニグラムモデル

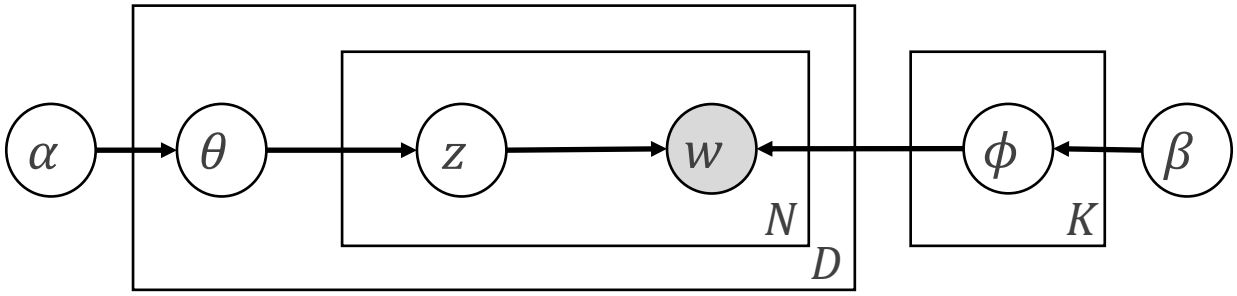


確率的潜在意味解析モデル
(PLSA)



潜在的ディリクレ配分法（LDA） [Blei et al., 2003]

単にトピックモデルと言えばこれを指すことも多いくらいに代表的なモデル

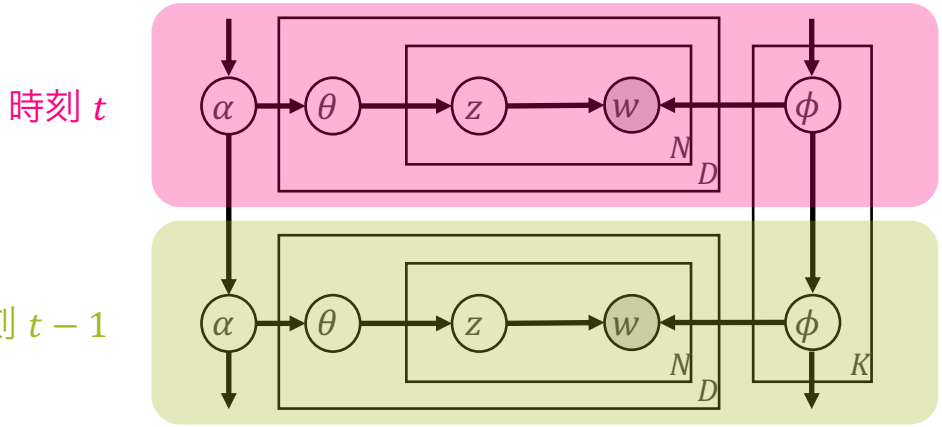


PLSAとの比較

- ハイパーパラメータ α と β が追加
- トピック分布と単語分布も確率的に生成
- オーバーフィット問題が軽減
- 未知文書も自然にモデル化できる
- 変分ベイズや（崩壊型）ギブスサンプリングを用いて推定

Dynamic Topic Model [Blei & Lafferty, 2006]

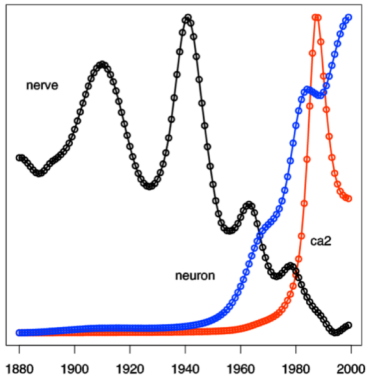
時間の経過にともなうトピックの変遷をモデル化



1881	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
brain	movement	brain	movement	movement	stimulate	record	respons	response	respons	cell	cell	neuron
movement	eye	eye	brain	sound	muscle	nerve	record	stimulate	cell	neuron	channel	active
action	right	movement	sound	muscle	sound	stimulate	stimulate	record	potential	response	neuron	brain
right	hand	right	nerve	active	movement	response	nerve	condition	stimul	active	ca2	cell
eye	brain	left	active	nerve	response	muscle	muscle	active	neuron	brain	active	fig
hand	left	hand	muscle	stimulate	nerve	electrode	active	potential	active	stimul	brain	response
left	action	nerve	left	fiber	frequency	active	frequency	stimulus	nerve	muscle	receptor	channel
muscle	muscle	vision	eye	reaction	fiber	brain	electrode	nerve	eye	system	muscle	receptor
nerve	sound	sound	right	brain	active	fiber	potential	subject	record	nerve	respons	synapse
sound	experiment	muscle	nervous	response	brain	potential	study	eye	abstract	receptor	current	signal

論文誌Scienceに掲載された論文のトピックの変遷

"Neuroscience"



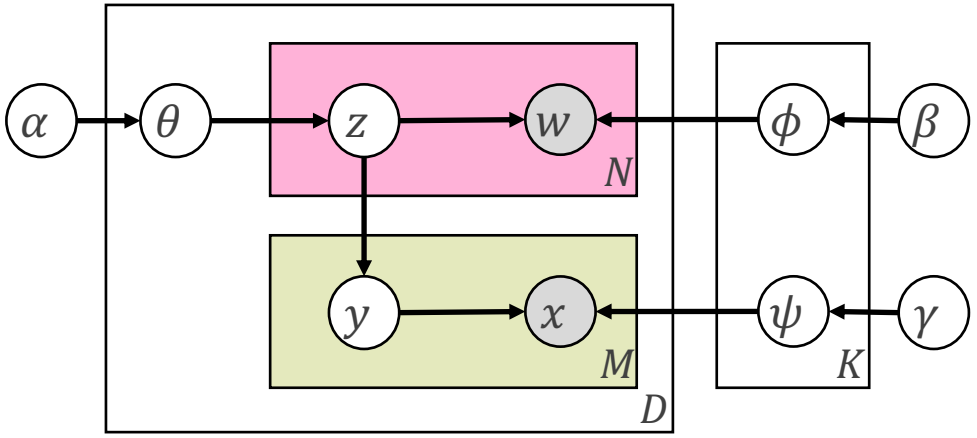
1887 Mental Science
1900 Hemianopsia in Migraine
1912 A Defence of the ``New Phrenology''
1921 The Synchronal Flashing of Fireflies
1932 Myoesthesis and Imageless Thought
1943 Acetylcholine and the Physiology of the Nervous System
1952 Brain Waves and Unit Discharge in Cerebral Cortex
1963 Errorless Discrimination Learning in the Pigeon
1974 Temporal Summation of Light by a Vertebrate Visual Receptor
1983 Hysteresis in the Force-Calcium Relation in Muscle
1993 GABA-Activated Chloride Channels in Secretory Nerve Endings

Correspondence Topic Model [Blei & Jordan, 2003]

主情報のトピックに依存する補助情報の生成も同時にモデル化

主情報（文書）

補助情報（タグ）



別の確率分布を使うことで画像・キャプション等もモデル化できる

未知画像

自動付与された
キャプション付与

True caption
market people
Corr–LDA
people market pattern textile display
GM–LDA
people tree light sky water
GM–Mixture
people market street costume temple

True caption
scotland water
Corr–LDA
scotland water flowers hills tree
GM–LDA
tree water people mountain sky
GM–Mixture
water sky clouds sunset scotland

True caption
bridge sky water
Corr–LDA
sky water buildings people mountain
GM–LDA
sky water people tree buildings
GM–Mixture
sky plane jet water snow

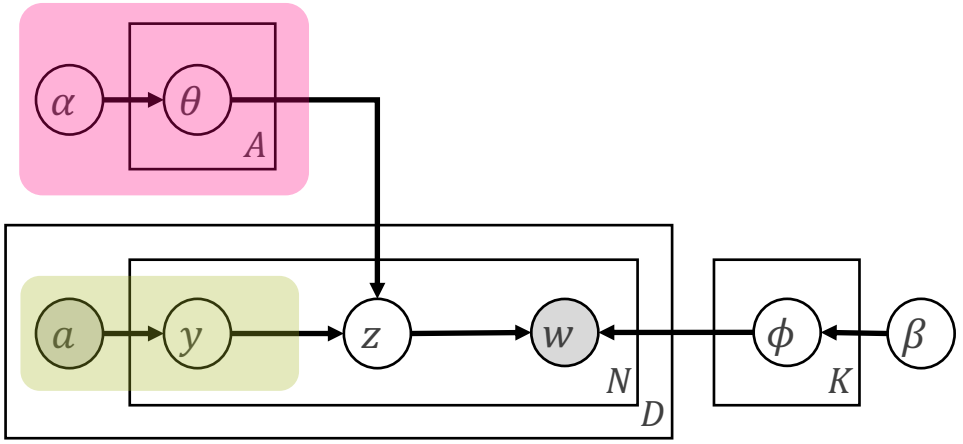
True caption
sky tree water
Corr–LDA
tree water sky people buildings
GM–LDA
sky tree fish water people
GM–Mixture
tree vegetables pumpkins water garden

Author Topic Model [Rosen-Zvi et al., 2004]

主情報のトピックが補助情報に依存する過程をモデル化

著者ごとに
トピック分布を生成

単語ごとに
著者を選択



**論文の各トピックにおける
頻出語・頻出著者**

TOPIC 10	
WORD	PROB.
SPEECH	0.1134
RECOGNITION	0.0349
WORD	0.0295
SPEAKER	0.0227
ACOUSTIC	0.0205
RATE	0.0134
SPOKEN	0.0132
SOUND	0.0127
TRAINING	0.0104
MUSIC	0.0102
AUTHOR	**PROB.**
Waibel_A	0.0156
Gauvain_J	0.0133
Lamel_L	0.0128
Woodland_P	0.0124
Ney_H	0.0080
Hansen_J	0.0078
Renals_S	0.0072
Noth_E	0.0071
Boves_L	0.0070
Young_S	0.0069

TOPIC 209	
WORD	PROB.
PROBABILISTIC	0.0778
BAYESIAN	0.0671
PROBABILITY	0.0532
CARLO	0.0309
MONTE	0.0308
DISTRIBUTION	0.0257
INFERENCE	0.0253
PROBABILITIES	0.0253
CONDITIONAL	0.0229
PRIOR	0.0219
AUTHOR	**PROB.**
Friedman_N	0.0094
Heckerman_D	0.0067
Ghahramani_Z	0.0062
Koller_D	0.0062
Jordan_M	0.0059
Neal_R	0.0055
Raftery_A	0.0054
Lukasiewicz_T	0.0053
Halpern_J	0.0052
Muller_P	0.0048

TOPIC 87	
WORD	PROB.
USER	0.2541
INTERFACE	0.1080
USERS	0.0788
INTERFACES	0.0433
GRAPHICAL	0.0392
INTERACTIVE	0.0354
INTERACTION	0.0261
VISUAL	0.0203
DISPLAY	0.0128
MANIPULATION	0.0099
AUTHOR	**PROB.**
Shneiderman_B	0.0060
Rauterberg_M	0.0031
Lavana_H	0.0024
Pentland_A	0.0021
Myers_B	0.0021
Minas_M	0.0021
Burnett_M	0.0021
Winiwarter_W	0.0020
Chang_S	0.0019
Korvemaker_B	0.0019

TOPIC 20	
WORD	PROB.
STARS	0.0164
OBSERVATIONS	0.0150
SOLAR	0.0150
MAGNETIC	0.0145
RAY	0.0144
EMISSION	0.0134
GALAXIES	0.0124
OBSERVED	0.0108
SUBJECT	0.0101
STAR	0.0087
AUTHOR	**PROB.**
Linsky_J	0.0143
Falcke_H	0.0131
Mursula_K	0.0089
Butler_R	0.0083
Bjorkman_K	0.0078
Knapp_G	0.0067
Kundu_M	0.0063
Christensen-J	0.0059
Cranmer_S	0.0055
Nagar_N	0.0050

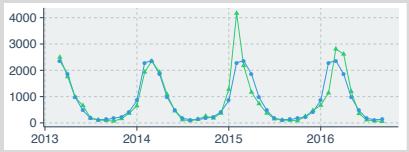
確率的投薬モデル

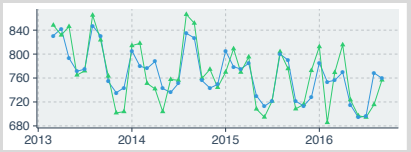
[Umemoto et al., 2019]

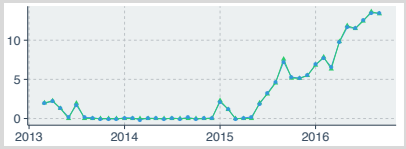
投薬トレンド

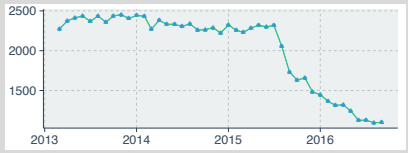
各疾患に対する医薬品の処方傾向は時間の経過とともに変化

- 日本では新医薬品の申請が年間100件以上
- 疾患・医薬品には季節性や流行といった多様な変化要因が存在









医薬品の供給を最適化したい

最新の処方知識を普及したい

医療費の変動を把握したい

全て（疾患・医薬品，変化の種類，…）を人手で定義・確認・調査するのは困難

多様に変化する投薬トレンドの自動解析

本研究のアプローチ

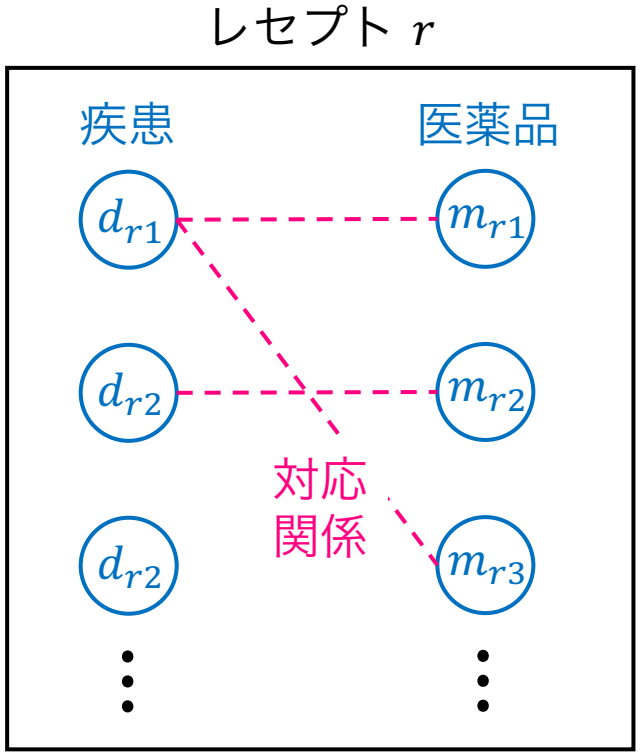
解析対象の医療情報として**レセプト（診療報酬明細書）**に着目

特徴

- **電子化:** 医療機関の93.2%, レセプトの98.2%
- **悉皆性:** 都道府県・全国レベルの医療状況を把握可能（国民皆保険制度）

網羅的なデータを計算機で分析可能

問題: レセプトでは対応関係が欠損



レセプトに含まれる

- 診断された疾患名
- 処方された医薬品

レセプトに含まれない

- 各疾患に対してどの医薬品が処方されたか（対応関係）

レセプトは月単位

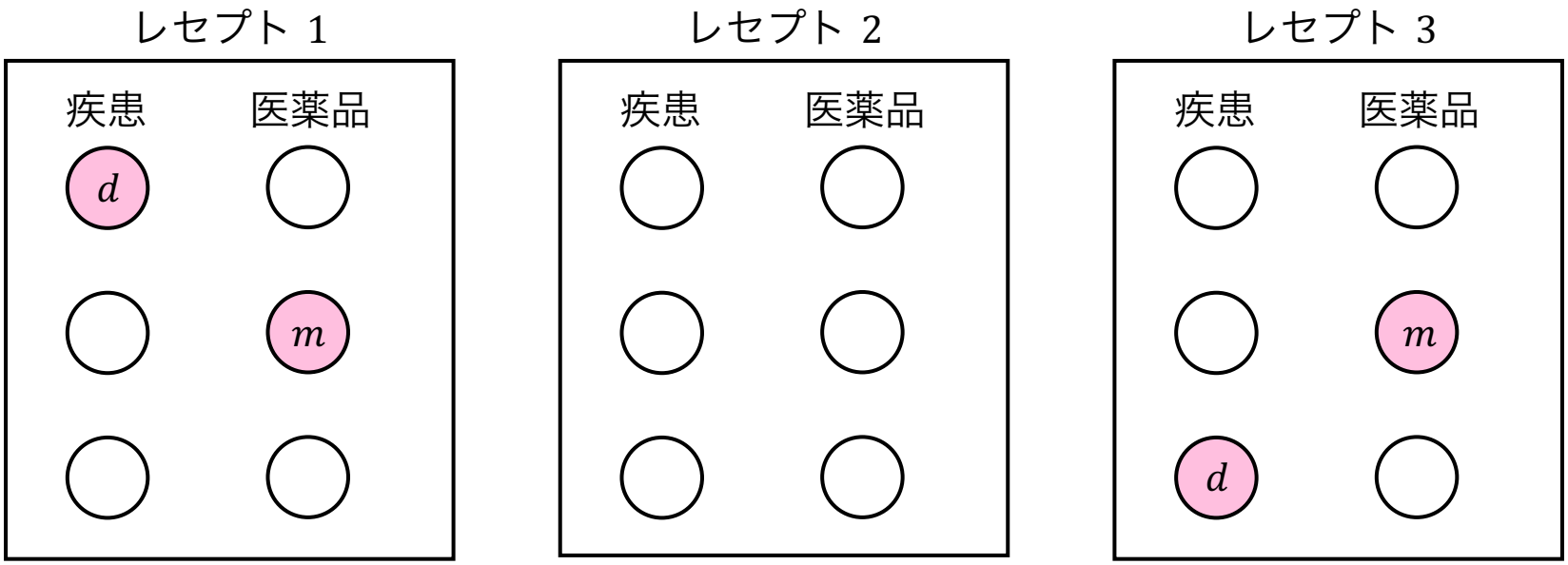
- 複数回の受診記録が同一のレセプトに集約

疾患･医薬品間の投薬回数を正確に復元するには？

レセプト中で欠損している**対応関係の推定**が必要

単純な解決法: 共起頻度に基づく関係推定

レセプト中での疾患•医薬品の共起頻度を処方回数をみなせば良いのでは？



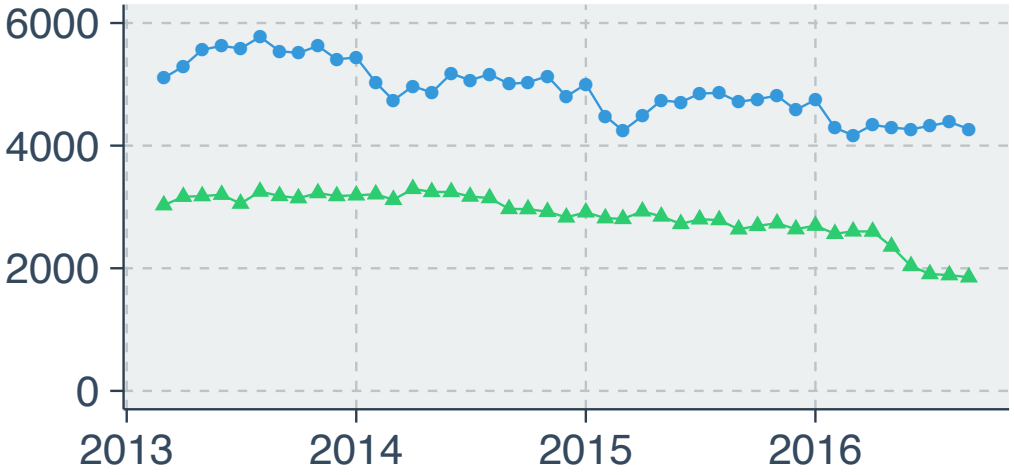
共起頻度に基づく疾患dと医薬品mの推定処方回数 ∝ 2回

単純な解決法: 共起頻度に基づく関係推定

レセプト中での疾患•医薬品の共起頻度を投薬回数をみなせば良いのでは？

共起頻度に基づき推定した**高血圧症**に対する投薬回数の時系列

経皮鎮痛消炎剤（効能なし） vs. **血圧降下剤（効能あり）**



共起頻度の問題点

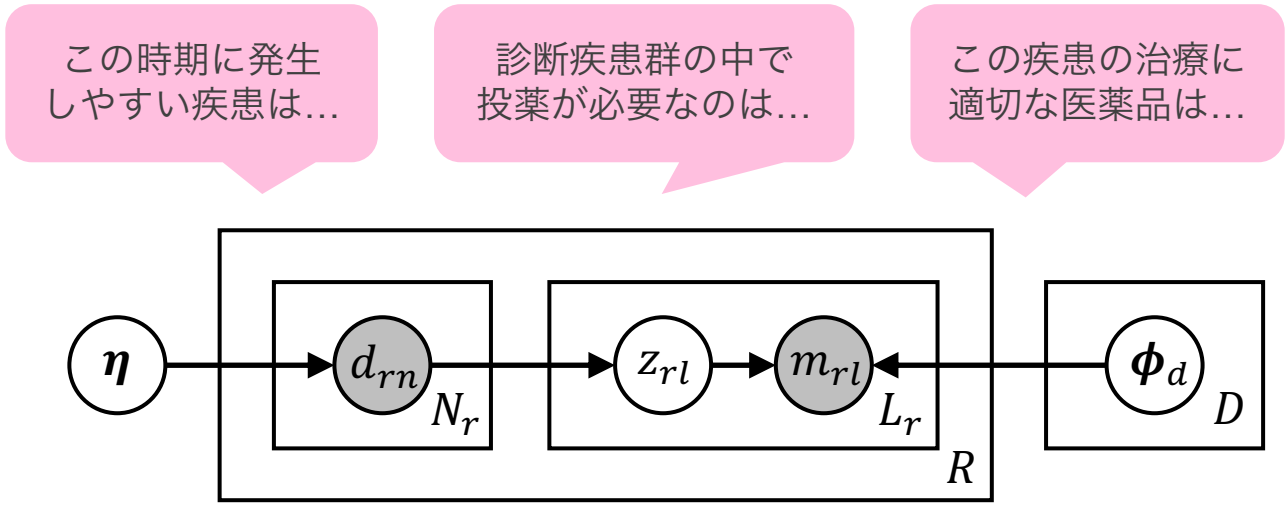
レセプト中で**出現頻度の高い医薬品**を
効能の有無にかかわらず**不当に高く評価**

提案: 確率的投薬モデル

医師の投薬行動
- **疾患の診断**
- **投薬対象の選択**
- **医薬品の処方**

を模倣しながら

レセプト中の疾患•医薬品を生成する潜在変数モデル

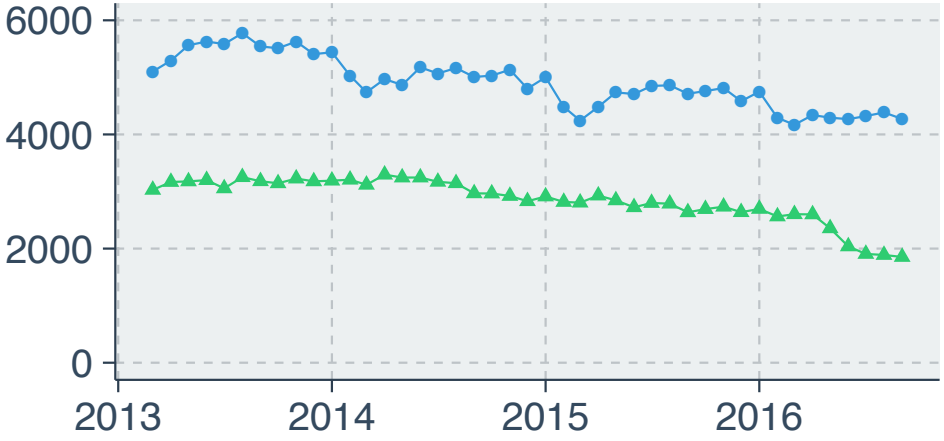


確率的投薬モデルによる関係推定

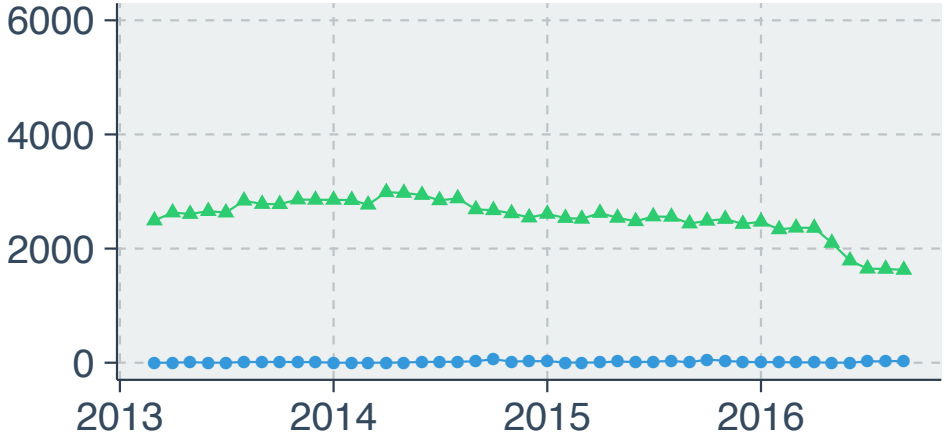
高血圧症に対する投薬回数の時系列

経皮鎮痛消炎剤（効能なし） vs. 血圧降下剤（効能あり）

共起頻度（ベースライン）



確率的投薬モデル（提案手法）

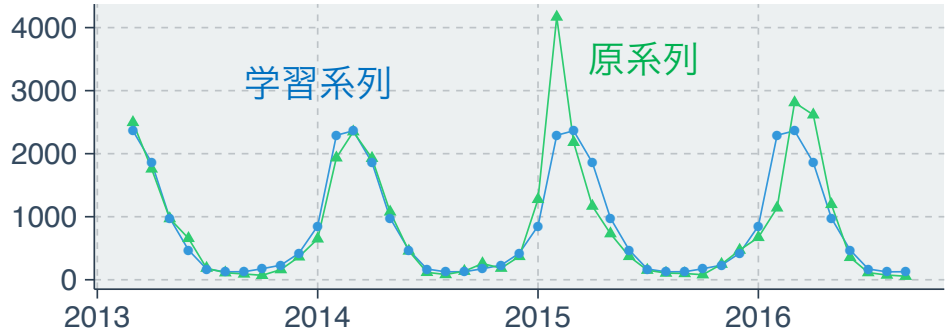


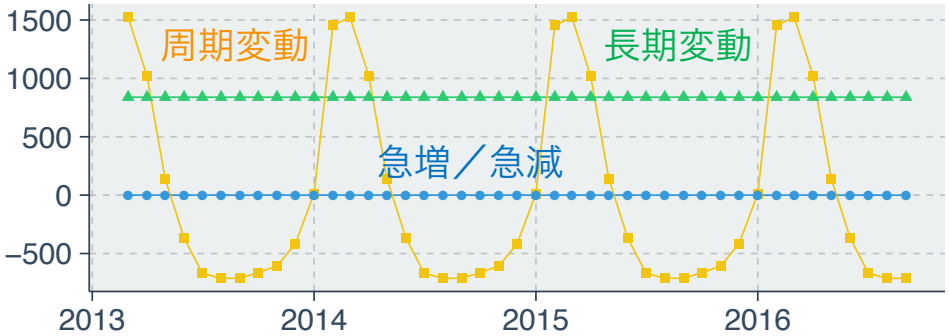
提案手法の利点

- 効能のない医薬品: **出現頻度が高くても投薬回数を低く推定**
- 効能のある医薬品: **推定数は元の頻度と同様の傾向を維持**

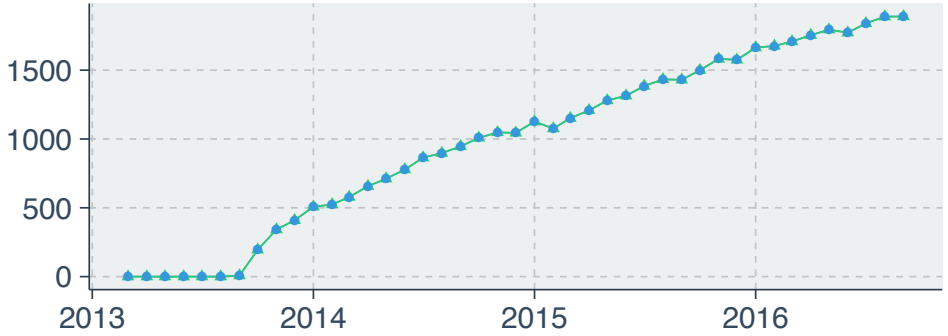
分析事例: 投薬トレンドの要因分解

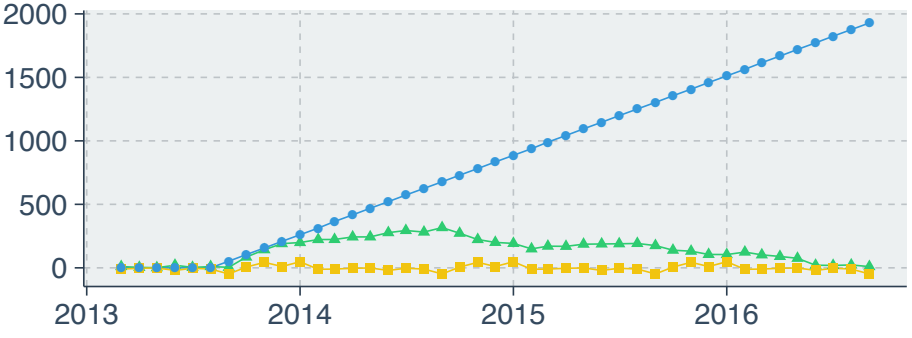
インフルエンザ（季節性）



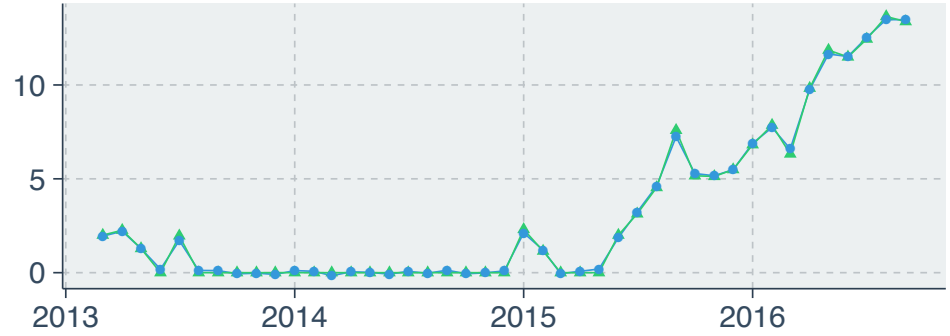


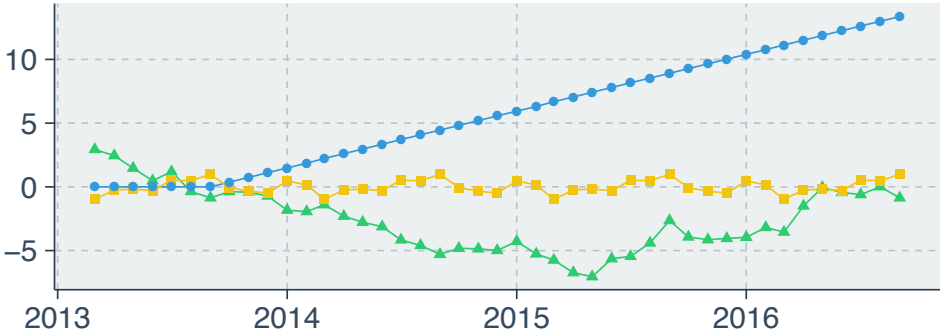
骨粗鬆症（新薬）





認知症（効能追加）

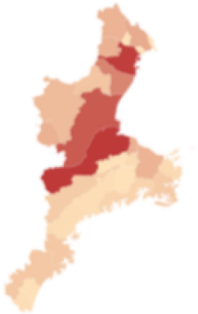




分析事例: 投薬トレンドの地域差の可視化

分解により急減／急増が検出された抗血小板薬の先発／後発医薬品

ジェネリック発売1か月前



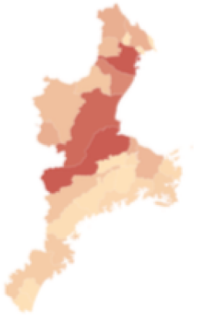
(a) Original

(b) Generic-1

(c) Generic-2

(d) Generic-3

ジェネリック発売1か月後



(e) Original

(f) Generic-1

(g) Generic-2

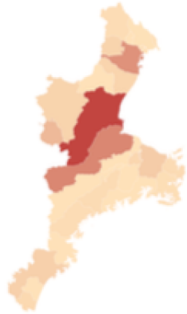
(h) Generic-3

ジェネリック発売1年後

(i) Original

(j) Generic-1

(k) Generic-2



(l) Generic-3

クリックモデル

検索結果ランキングのための様々な要因 [1, 2]

On-The-Page Factors

These elements are in the direct control of the publisher

Content

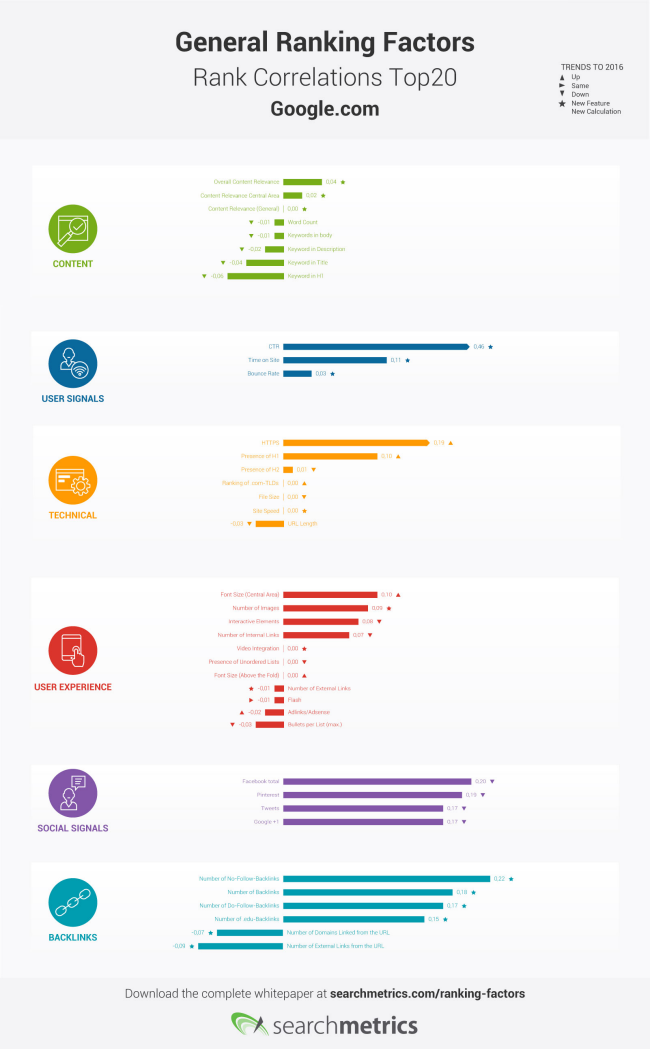
Cq	QUALITY	Are pages well written & have substantial quality content?
Cr	RESEARCH	Have you researched the keywords people may use to find your content?
Cw	WORDS	Do pages use words & phrases you hope they'll be found for?
Cf	FRESH	Are pages fresh & about "hot" topics?
Cv	VERTICAL	Do you have image, local, news, video or other vertical content?
Ca	ANSWERS	Is your content turned into direct answers within search results?
Vt	THIN	Is content "thin" or "shallow" & lacking substance?

Architecture

Ac	CRAWL	Can search engines easily "crawl" pages on site?
Am	MOBILE	Does your site work well for mobile devices?
Ad	DUPLICATE	Does site manage duplicate content issues well?
As	SPEED	Does site load quickly?
Au	URLS	Do URLs contain meaningful keywords to page topics?
Ah	HTTPS	Does site use HTTPS to provide secure connection for visitors?
Vc	CLOAKING	Do you show search engines different pages than humans?

HTML

Ht	TITLES	Do HTML title tags contain keywords relevant to page topics?
Hd	DESCRIPTION	Do meta description tags describe what pages are about?
Hs	STRUCTURE	Do pages use structured data to enhance listings?
Hh	HEADERS	Do headlines & subheads use header tags with relevant keywords?
Vs	STUFFING	Do you excessively use words you want pages to be found for?
Vh	HIDDEN	Do colors or design "hide" words you want pages to be found for?



Off-The-Page Factors

Elements influenced by readers, visitors & other publishers

Trust

Ta	AUTHORITY	Do links, shares & other factors make pages trusted authorities?
Te	ENGAGE	Do visitors spend time reading or "bounce" away quickly?
Th	HISTORY	Has site or its domain been around a long time, operating in same way?
Vd	PIRACY	Has site been flagged for hosting pirated content?
Va	ADS	Is content ad-heavy? Do you make use of intrusive interstitials?

Links

Lq	QUALITY	Are links from trusted, quality or respected web sites?
Lt	TEXT	Do links pointing at pages use words you hope they'll be found for?
Ln	NUMBER	Do many links point at your web pages?
Vp	PAID	Have you purchased links in hopes of better rankings?
Vl	SPAM	Have you created links by spamming blogs, forums or other places?

Personal

Pc	COUNTRY	What country is someone located in?
Pl	LOCALITY	What city or local area is someone located in?
Ph	HISTORY	Has someone regularly visited your site?

Social

Sr	REPUTATION	Do those respected on social networks share your content?
Ss	SHARES	Do many share your content on social networks?

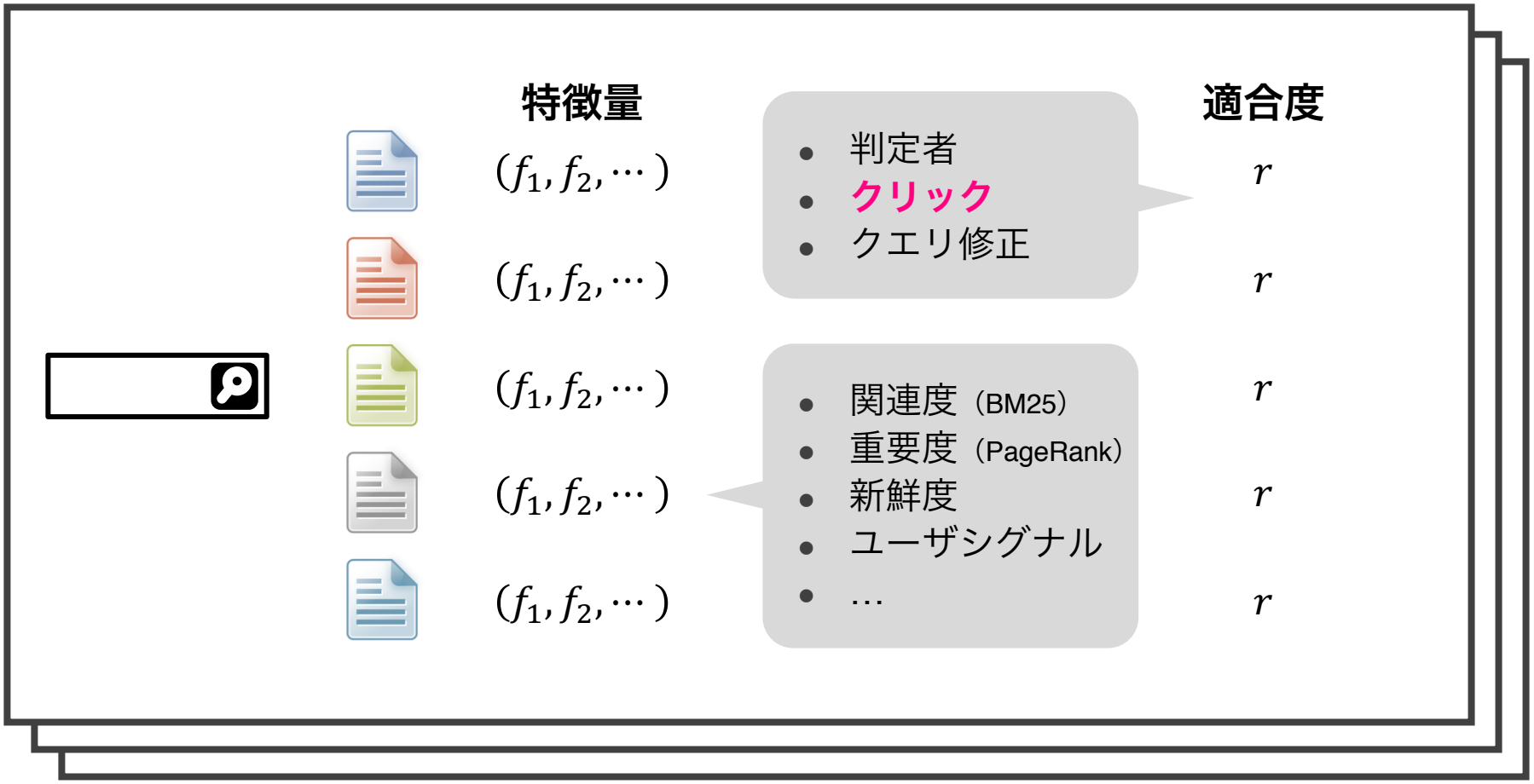
WRITTEN BY: Search Engine Land

CREATED BY: COLUMN FIVE

LEARN MORE: http://selnd.com/seotable

[1] https://searchengineland.com/seotable
[2] https://www.shoutmeloud.com/ranking-factors

ランキング学習におけるクリックの重要性



適合度関数を学習して未知クエリでもランキングを生成

クリックモデル

Web検索結果ページ上でのユーザのクリック行動を記述する確率モデル

単純なモデル｜Global CTR Model

クリック確率が全ての文書で共通だと仮定

検索クエリ

$p(C_{d_1} = 1) = \rho$

$p(C_{d_2} = 1) = \rho$

$p(C_{d_3} = 1) = \rho$

$p(C_{d_4} = 1) = \rho$

$p(C_{d_5} = 1) = \rho$

$$\rho = \frac{\#\{\text{clicks}\}}{\#\{\text{shown doc}\}}$$

位置バイアス [Joachims et al., 2007]

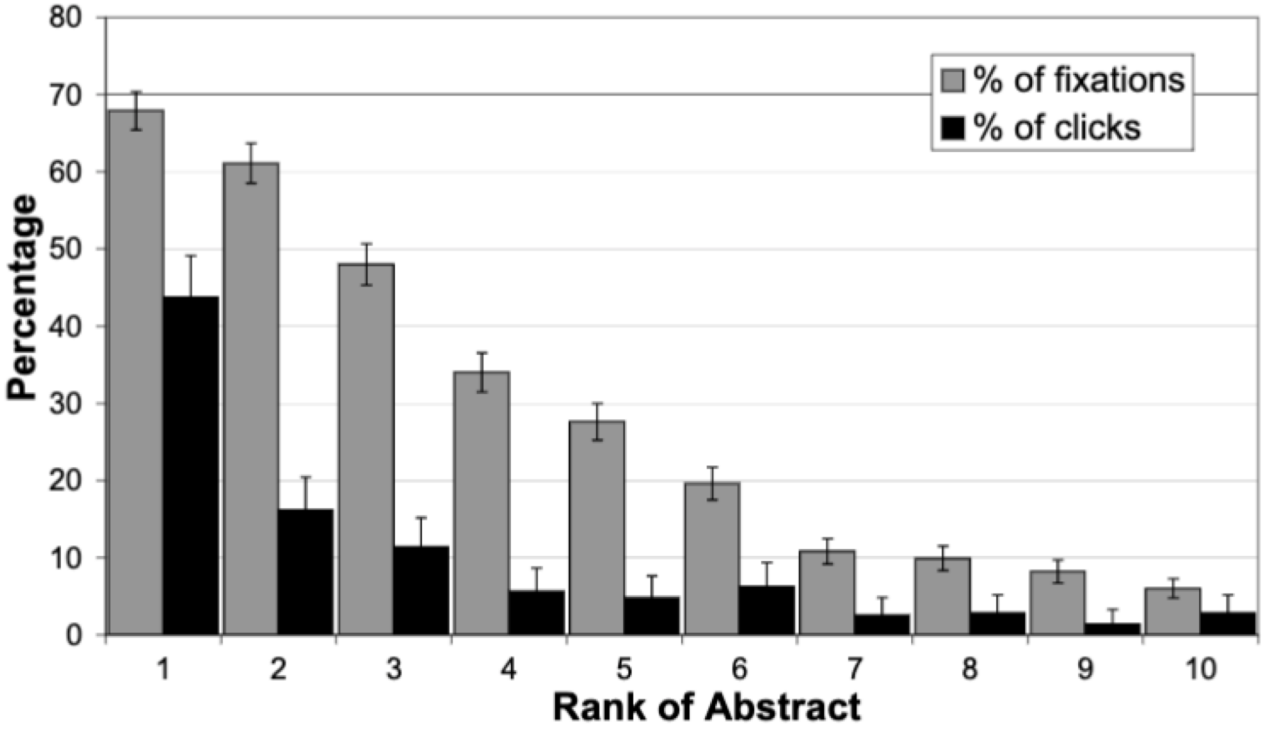
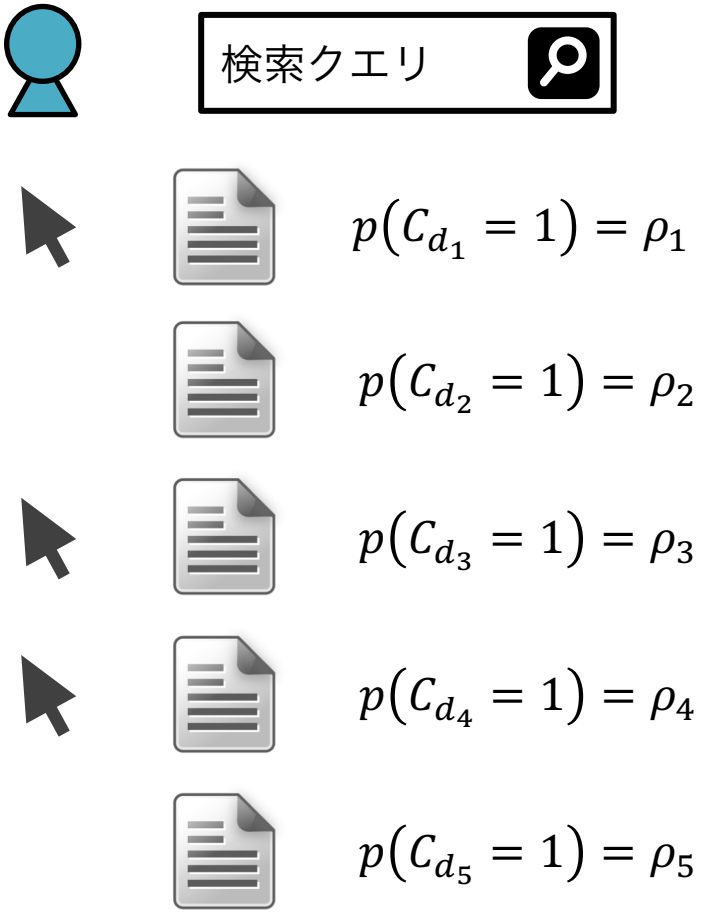


Fig. 1. Percentage of times an abstract was viewed/clicked depending on the rank of the result.

単純なモデル｜Rank-based CTR Model

クリック確率が文書の位置によって決まると仮定



$$\rho_r = \frac{\#\{\text{clicks at rank } r\}}{\#\{\text{shown docs at rank } r\}}$$

単純なモデル｜Query-document CTR Model

クリック確率がクエリと文書のペアごとに異なると仮定

検索クエリ q

$$p(C_{d_1} = 1) = \rho_{qd_1}$$

$$p(C_{d_2} = 1) = \rho_{qd_2}$$

$$p(C_{d_3} = 1) = \rho_{qd_3}$$

$$p(C_{d_4} = 1) = \rho_{qd_4}$$

$$p(C_{d_5} = 1) = \rho_{qd_5}$$

$$\rho_{qd} = \frac{\#\{d \text{ is clicked for } q\}}{\#\{d \text{ is shown for } q\}}$$

Position-based Model

クリック確率を調査（Examination）と魅力（Attractiveness）に分けてモデル化

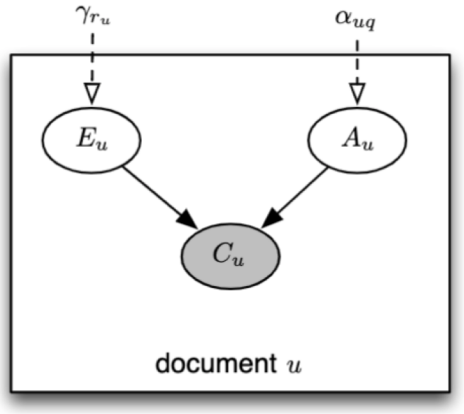
- 調査: スニペット（検索結果の要約文）を読むこと（←位置に依存）
- 魅力: スニペット調査後にクリックしたくなること（←クエリ・文書ペアに依存）

検索クエリ q

$$p(E_1 = 1) = \gamma_1, \qquad p(A_{d_1} = 1) = \alpha_{qd_1}$$

$$p(E_2 = 1) = \gamma_2, \qquad p(A_{d_2} = 1) = \alpha_{qd_2}$$

$$p(E_3 = 1) = \gamma_3, \qquad p(A_{d_3} = 1) = \alpha_{qd_3}$$



$$p(C_d = 1) = p(E_{r_d} = 1) \cdot p(A_d = 1) = \gamma_{r_d} \cdot \alpha_{qd}$$

Cascade Model

ユーザが検索結果を上から下に調査する行動をモデル化

- 現在の位置の文書をクリックしたらそこで調査を終了
- それ以外の場合は以降の文書の調査を続ける

$$E_r = 1 \ \wedge \ A_{d_r} = 1 \Leftrightarrow C_r = 1$$

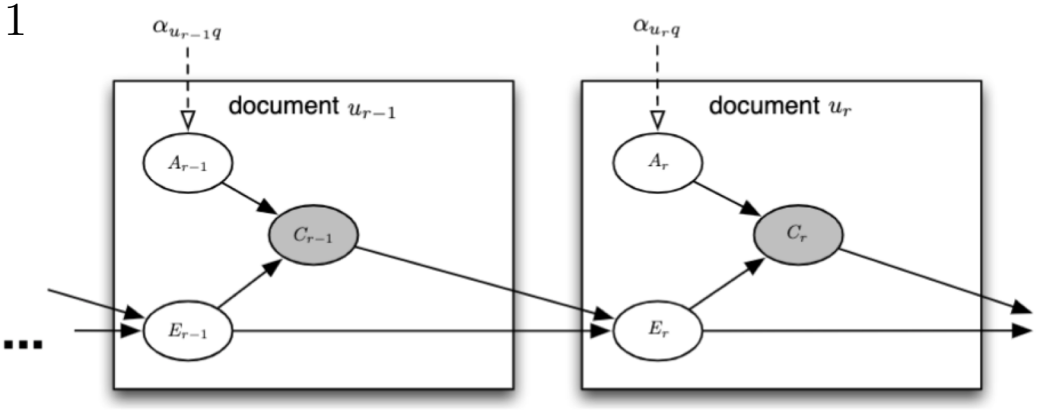
$$P(A_{d_r} = 1) = \alpha_{qd_r}$$

$$P(E_1 = 1) = 1$$

$$P(E_r = 1 | E_{r-1} = 0) = 0$$

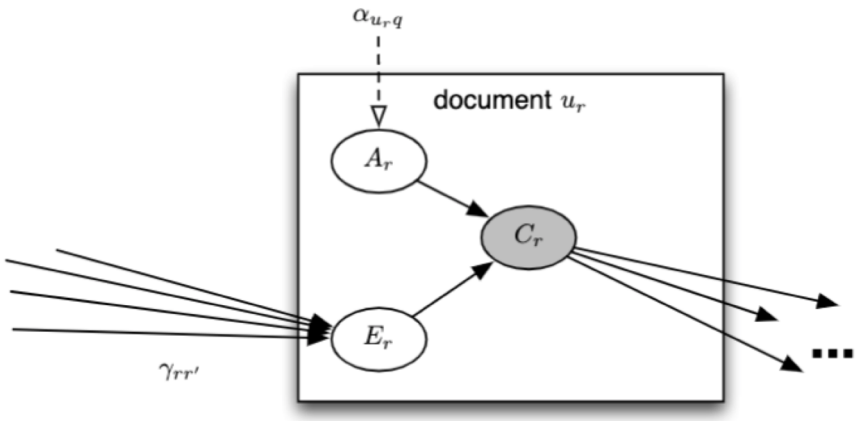
$$P(E_r = 1 | C_{r-1} = 1) = 0$$

$$P(E_r = 1 | E_{r-1} = 1, C_{r-1} = 0) = 1$$



更なる拡張

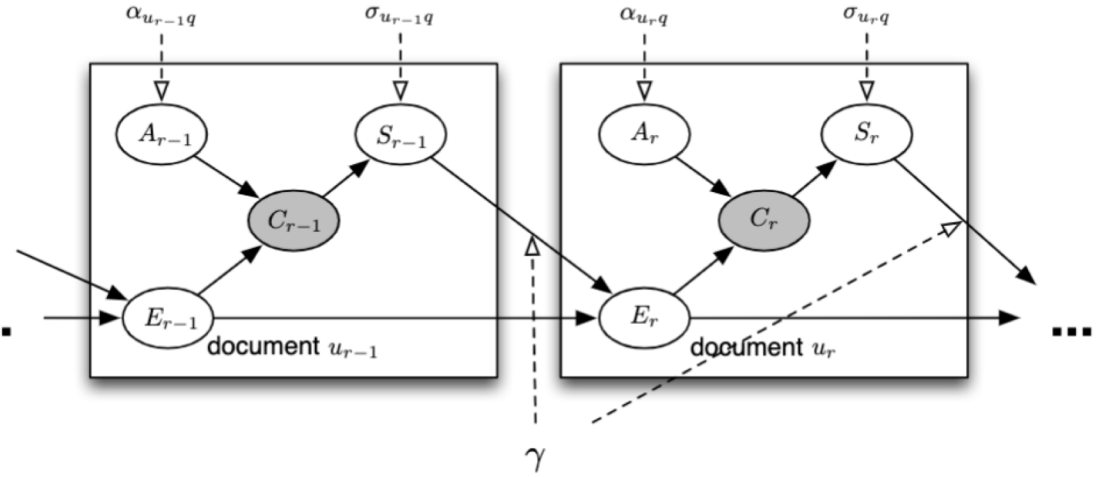
User Browsing Model



- Position-based Modelの拡張
- 最後のクリック文書と
  現在の文書との関係を考慮

$$P(E_r = 1 | C_{r'} = 1, C_{r'+1} = 0, \ldots, C_{r-1} = 0) = \gamma_{rr'}$$

Dynamic Bayesian Network Model



- Cascade Modelの拡張
- クリックでなく満足した時に
  調査を終了する

$$P(S_r = 1 | C_r = 1) = \sigma_{q d_r}$$

性能比較 [Grotov et al., 2015]

全ての評価指標に対して最良な万能モデルはこの中にはない模様...

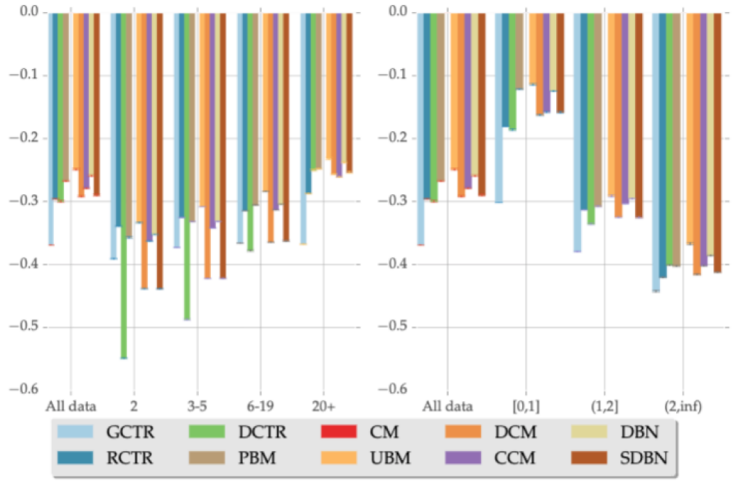


Fig. 1: Log-likelihood of click models, grouped by query frequency (left) and click entropy (right).

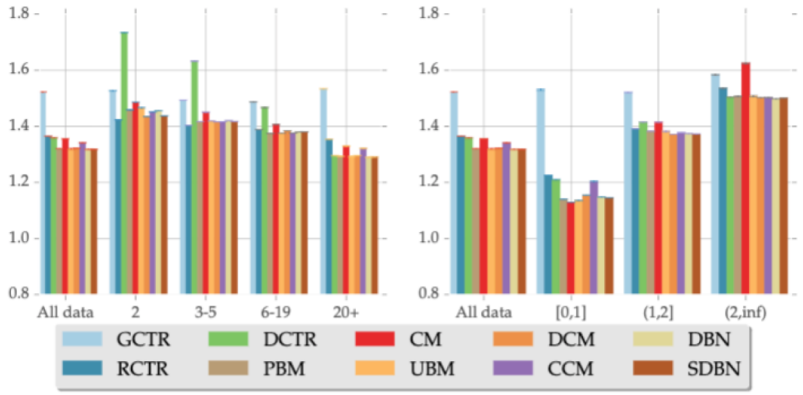


Fig. 2: Perplexity of click models, grouped by query frequency (left) and click entropy (right).

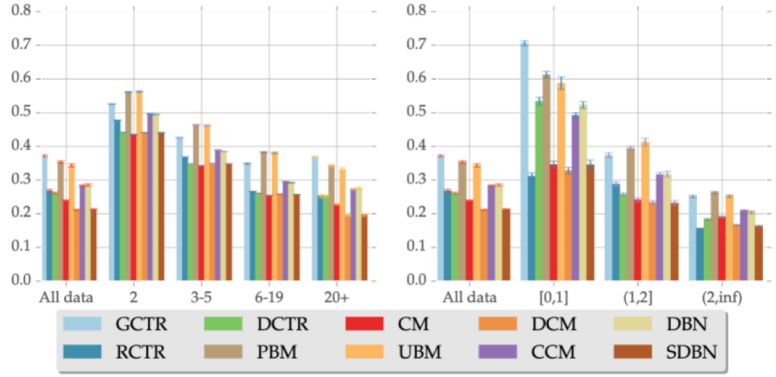


Fig. 3: Click-through rate prediction RMSE of click models, grouped by query frequency (left) and click entropy (right).

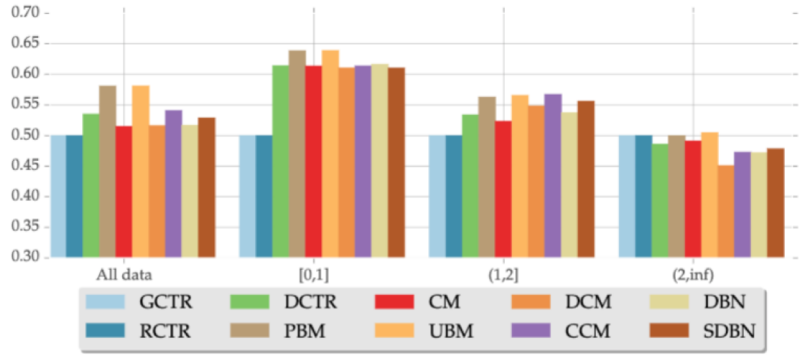


Fig. 4: Relevance prediction of click models on click entropy

まとめ

Q. 観測データの背後に隠れた情報を推定するには？

A. 潜在変数モデルを使ってみよう

- 潜在変数を用いてデータの生成過程を統計にモデル化
- 確率分布の決定→教師なし学習→潜在変数の推定値の応用
- 😆 言語・画像など多岐にわたる問題に適用可能
- 😢 推論部分の導出がちょっとつらい

参考文献

- Shlens, J. (2014). A tutorial on principal component analysis. arXiv preprint arXiv:1404.1100.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.
- Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In Proceedings of the 23rd international conference on Machine learning (pp. 113-120). ACM.
- Blei, D. M., & Jordan, M. I. (2003). Modeling annotated data. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval (pp. 127-134). ACM.
- Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. In Proceedings of the 20th conference on Uncertainty in artificial intelligence (pp. 487-494). AUAI Press.
- Umemoto, K., Goda, K., Mitsutake, N., & Kitsuregawa, M. (2019). A Prescription Trend Analysis using Medical Insurance Claim Big Data. In 2019 IEEE 35th International Conference on Data Engineering (ICDE) (pp. 1928-1939). IEEE.
- Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, F., & Gay, G. (2007). Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. ACM Transactions on Information Systems (TOIS), 25(2), 7.
- Grotov, A., Chuklin, A., Markov, I., Stout, L., Xumara, F., & de Rijke, M. (2015). A comparative study of click models for web search. In International Conference of the Cross-Language Evaluation Forum for European Languages (pp. 78-90). Springer, Cham.