潜在トピックモデルに基づくデータマイニング

NTTコミュニケーション科学基礎研究所
岩田具治

トピックモデルとは

- 文書が生成される過程を確率的に表現したモデル
- 様々な離散データで有効性が確認
  - 文書、購買、ネットワーク、画像、音楽
- 幅広い応用範囲
  - 情報検索、可視化、画像認識、推薦システム、音声認識
- 拡張が容易
- 実装が簡単

トピック抽出

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

入力:
文書集合

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

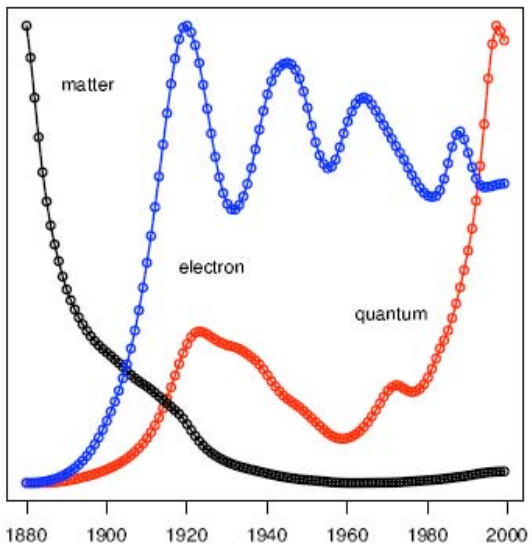
Figure 8: An example article from the AP corpus. Each color codes a different factor from which the word is putatively generated.

D. Blei, A. Ng, M. Jordan, Latent Dirichlet Allocation, JMLR2003

テキスト時間発展

1881	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
force	motion	magnet	force	atom	ray	energy	energy	radiat	electron	electron	electron	state
energy	force	electric	magnet	theory	measure	measure	radiat	energy	energy	energy	atom	energy
motion	magnet	measure	theory	electron	energy	electron	ray	electron	atom	particle	energy	electron
differ	energy	force	electric	energy	theory	light	electron	measure	measure	field	structur	magnet
light	measure	theory	atom	measure	light	atom	measure	ray	radiat	radiat	field	field
measure	differ	system	system	ray	wave	particle	atom	atom	field	model	model	atom
magnet	direct	motion	measure	electr	radiat	ray	particle	field	ray	atom	state	system
direct	line	line	line	line	atom	radiat	two	two	model	two	two	two
matter	result	point	energy	force	electric	point	light	particle	particle	ray	magnet	quantum
result	light	differ	body	value	value	theory	absorpt	observe	magnet	measure	ray	physic

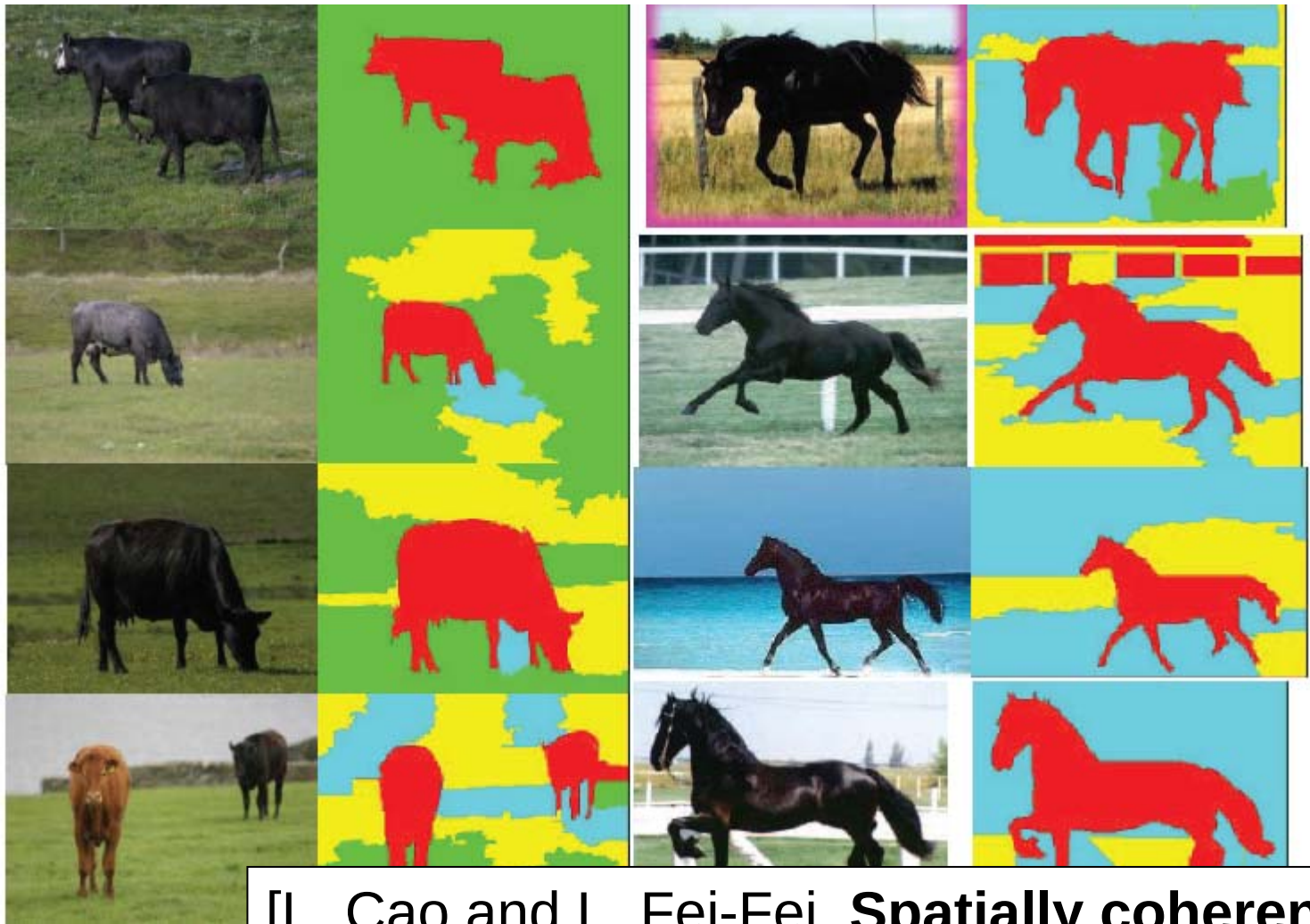
"Atomic Physics"



1881 On Matter as a form of Energy
1892 Non-Euclidean Geometry
1900 On Kathode Rays and Some Related Phenomena
1917 ``Keep Your Eye on the Ball"
1920 The Arrangement of Atoms in Some Common Metals
1933 Studies in Nuclear Physics
1943 Aristotle, Newton, Einstein. II
1950 Instrumentation for Radioactivity
1965 Lasers
1975 Particle Physics: Evidence for Magnetic Monopole Obtained
1985 Fermilab Tests its Antiproton Factory
1999 Quantum Computing with Electrons Floating on Liquid Helium

D. Blei, J. Lafferty, Dynamic Topic Models, ICML2006

画像認識



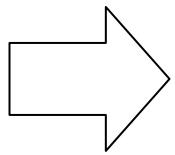
[L. Cao and L. Fei-Fei. **Spatially coherent latent topic model for concurrent object segmentation and classification . ICCV2007]**

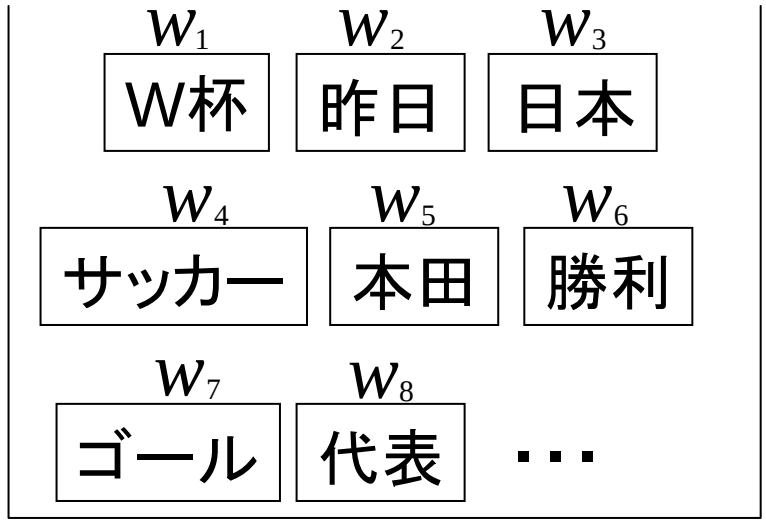
目次

- トピックモデルの説明
  - 多項分布、混合多項分布
  - PLSA、LDA
  - グラフィカルモデル
- トピックモデルに関する研究紹介
  - 購買行動解析のためのトピック追跡モデル

bag-of-word

昨日，サッカーW杯にて日本代表が本田のゴールにより勝利・・・

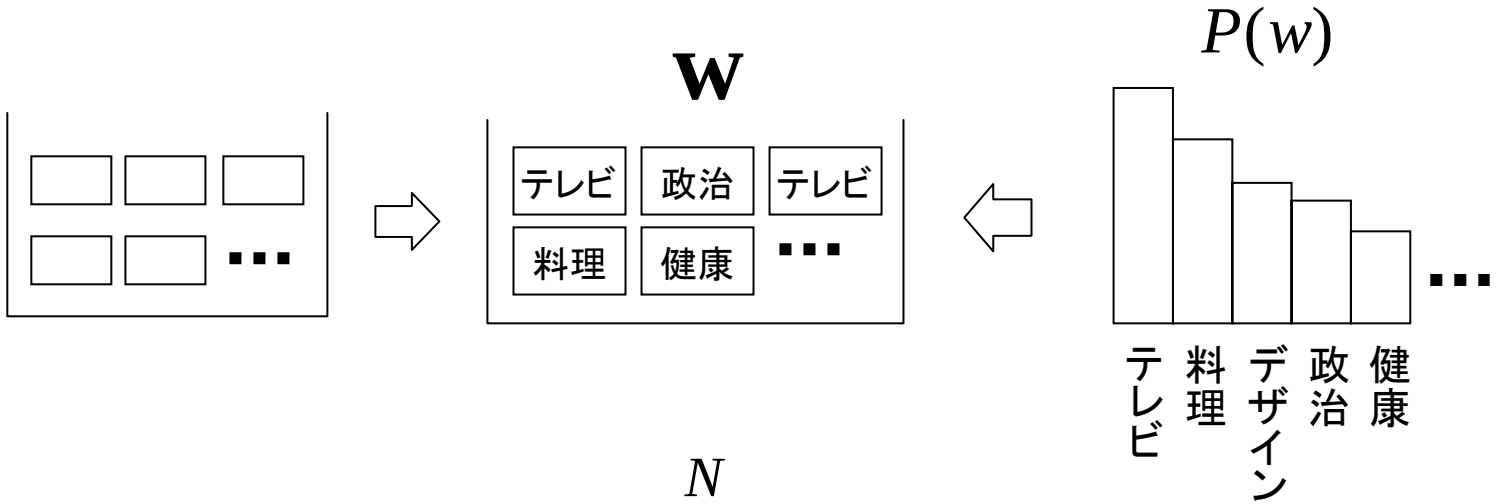




w

- 文書を単語集合として表現（順序なし）
- 情報検索、文書分類などで用いられる

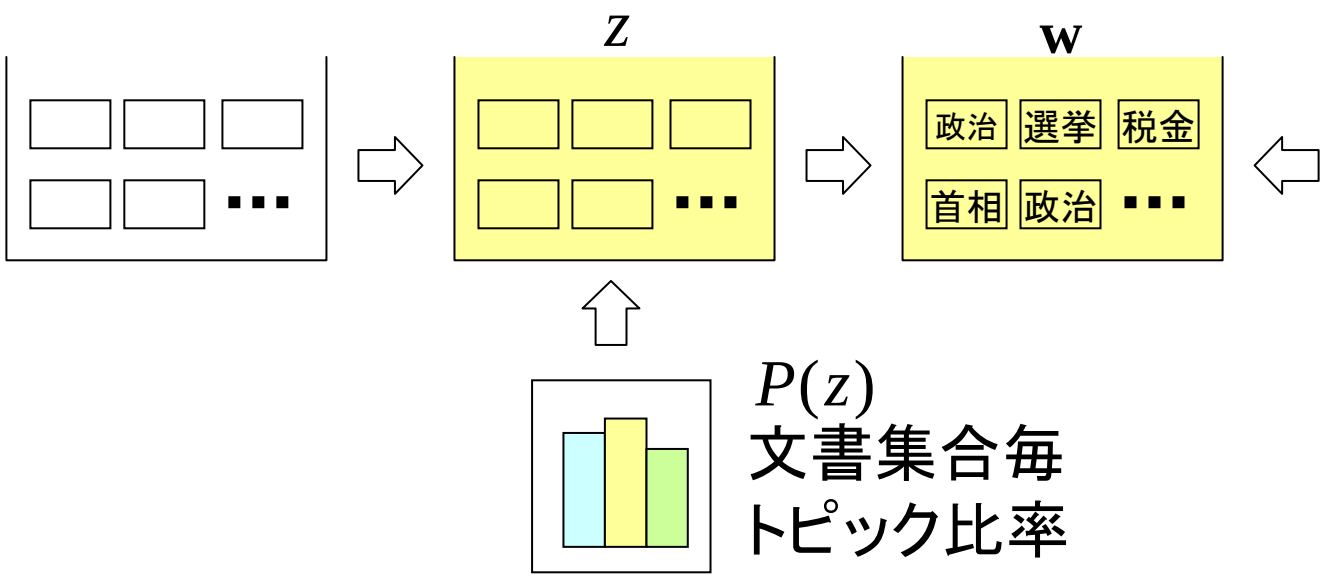
多項分布



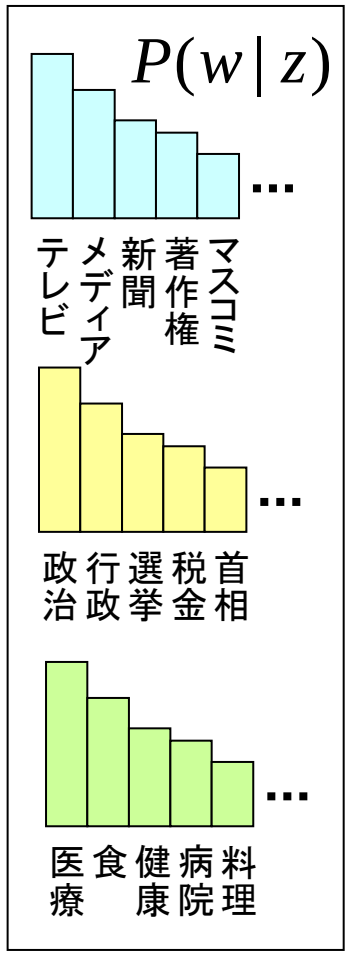
$$P(\mathbf{w}) = \prod_{n=1}^{N} P(w_n)$$

- 全文書の単語が同一の分布から生成

混合多項分布



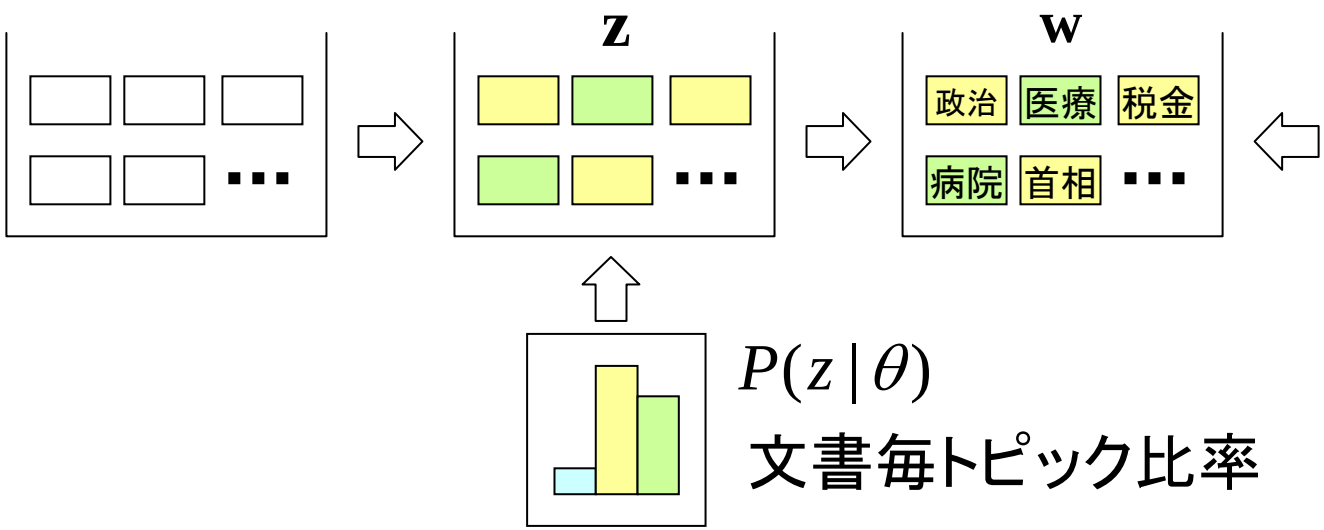
$P(z)$
文書集合毎
トピック比率

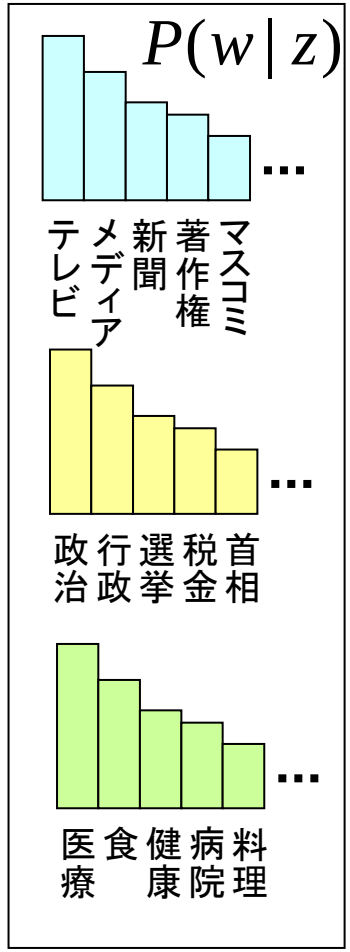


$$P(\mathbf{w}) = \sum_{z} P(z) \prod_{n=1}^{N} P(w_n \mid z)$$

- 文書毎にトピックを選択
- 1文書の単語が同一の分布から生成

トピックモデル





$$P(\mathbf{w}) = \prod_{n=1}^{N} \sum_{z} P(z|\theta) P(w_n|z)$$

- 単語毎にトピックを選択
- １文書の単語が複数の分布から生成

PLSAとLDA

- 代表的トピックモデル
- PLSA（=PLSI）: Probabilistic Latent Semantic Analysis

$$P(\mathbf{w}) = \prod_{n=1}^{N} \sum_{z} P(z \mid \theta) P(w_n \mid z)$$

- LDA: Latent Dirichlet Allocation
  - PLSA+トピック比率にディリクレ事前分布を仮定

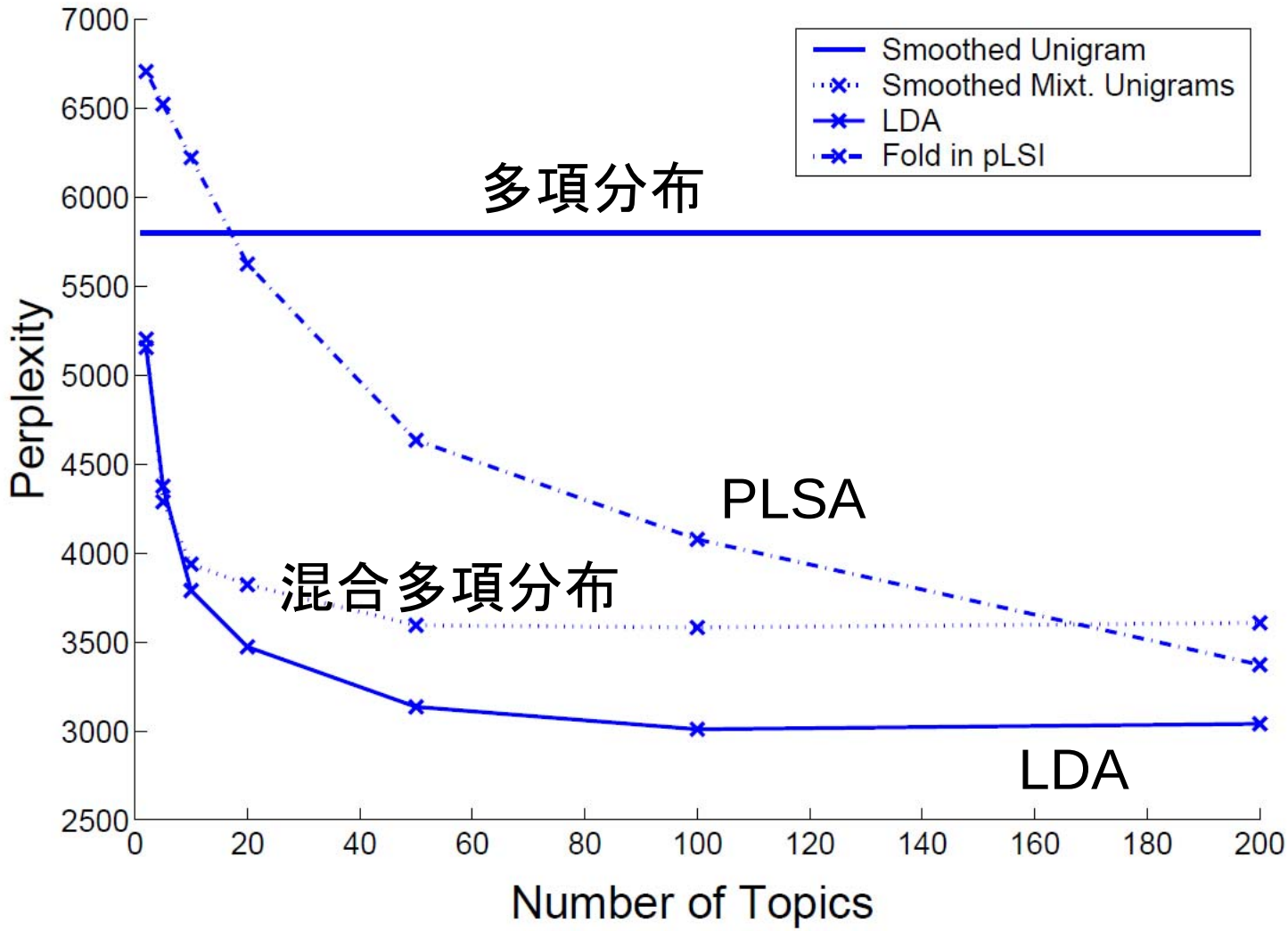
$$P(\mathbf{w}) = \int P(\theta) \prod_{n=1}^{N} \sum_{z} P(z \mid \theta) P(w_n \mid z) d\theta$$

ディリクレ事前分布

トピック比率

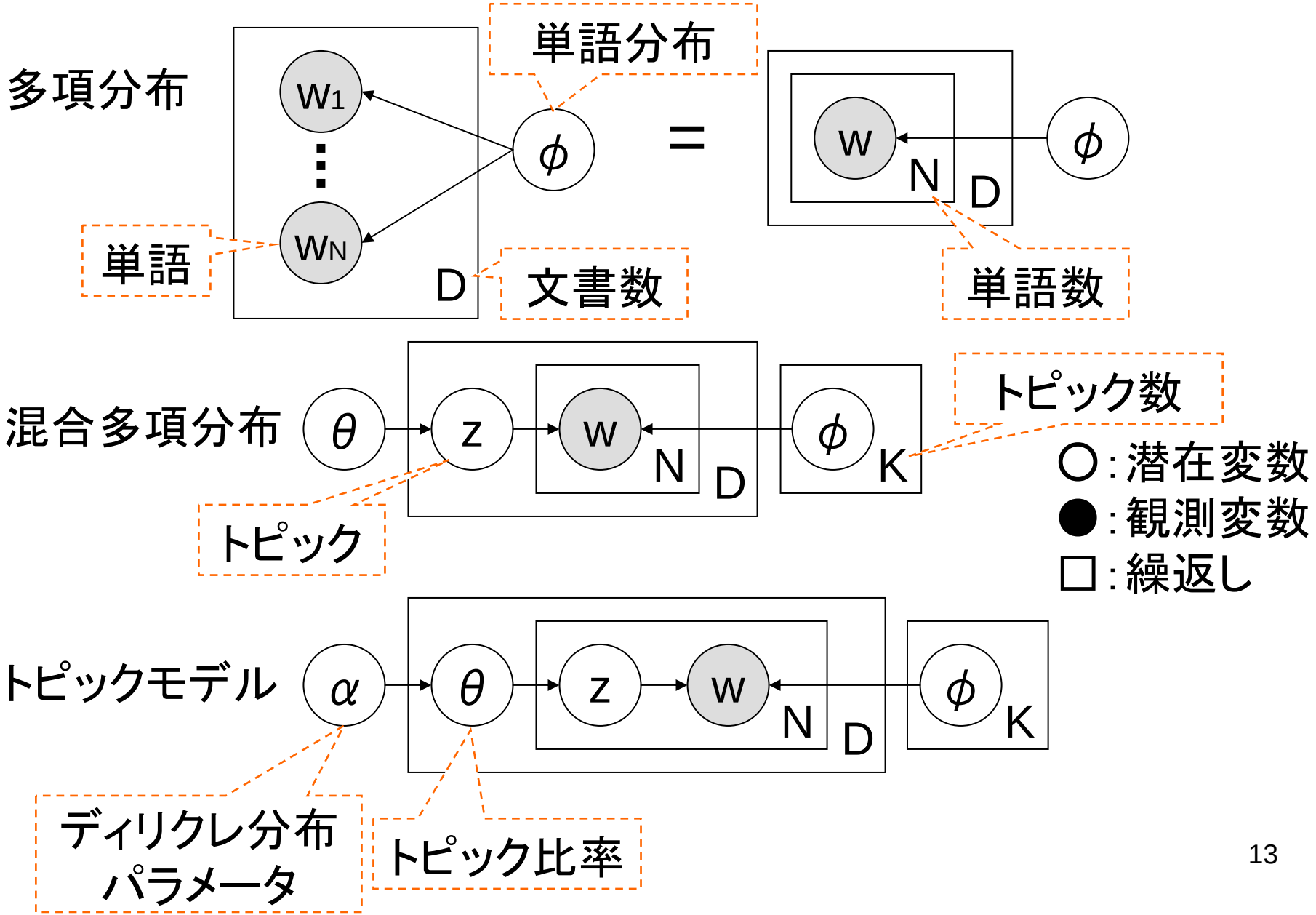
単語分布

文書データを用いた比較実験



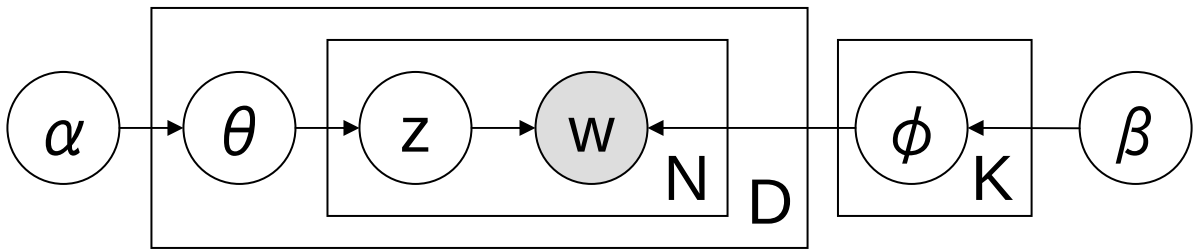
D. Blei, A. Ng, M. Jordan, Latent Dirichlet Allocation, JMLR2003

グラフィカルモデル


多項分布
単語分布
w_1
w_N
ϕ
=
w
N
D
単語
文書数
単語数
混合多項分布
θ
z
トピック数
K
トピック
○：潜在変数
●：観測変数
□：繰返し
トピックモデル
α
ディリクレ分布
パラメータ
トピック比率

学習

- Collapsedギブスサンプリング
  - θ と ϕ を積分消去してギブスサンプリング



$$P(z_n = k \mid \mathbf{W}, \mathbf{Z}_{\backslash n}) \propto \left(N_{dk\backslash n} + \alpha\right) \cdot \frac{N_{kw_n\backslash n} + \beta}{N_{k\backslash n} + \beta V}$$

$\mathbf{w}, \mathbf{z}_{\backslash n}$

政治 医療 税金

病院 首相 ▪▪▪

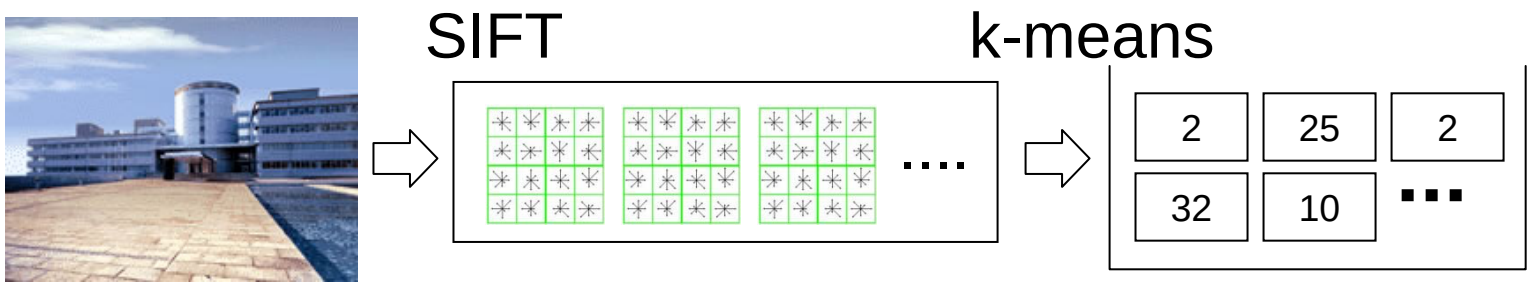
文書dのなかの
トピックkの数

トピックkのなかの
単語w_nの割合

（n番目の単語を除いたときの）

応用

- 画像認識
  - 文書=写真、単語=visual word

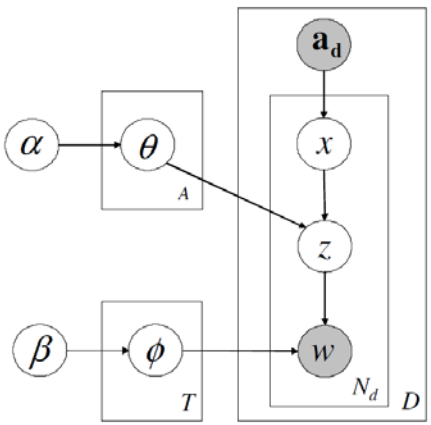


- 推薦システム
  - 文書=ユーザ，単語=商品
- ネットワーク、音楽、株価・・・

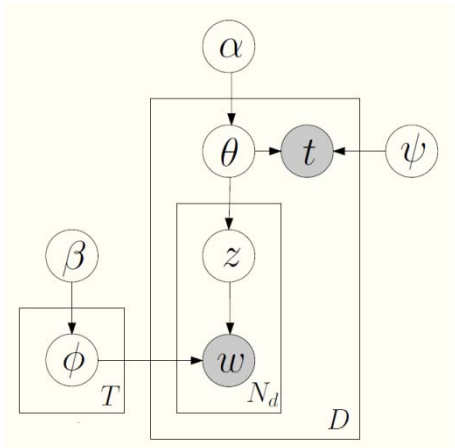
拡張

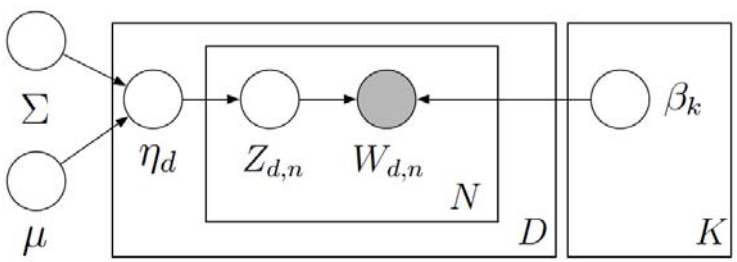
- 観測データの種類を増やす（著者、時間・・・）
- トピックの性質を変える（相関、階層・・・）

著者 [Rosen-Zvi, et al. 04]

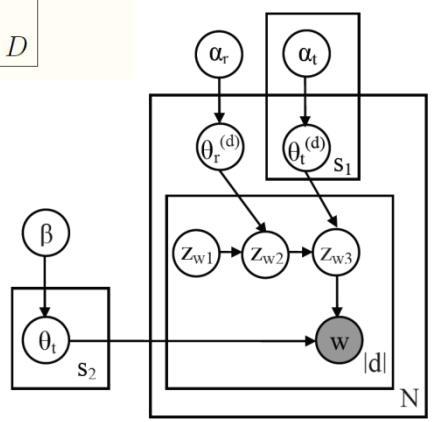


時間 [Wang and McCallum, 06]





相関 [Blei and Lafferty, 07]



階層 [Li and McCallum, 06]

トピックモデルに関連する成果

- 可視化
  - T. Iwata, T. Yamada, N. Ueda, “Probabilistic Latent Semantic Visualization: Topic Model for Visualizing Documents,” KDD2008
- 購買行動解析
  - T. Iwata, S. Watanabe, T. Yamada, N. Ueda, "Topic Tracking Model for Analyzing Consumer Purchase Behavior,“ IJCAI2009
- ソーシャルアノテーション解析
  - T. Iwata, T. Yamada, N. Ueda, "Modeling Social Annotation Data with Content Relevance using a Topic Model," NIPS2009
- 多重スケール時間発展解析
  - T. Iwata, T. Yamada, Y. Sakurai, N. Ueda, "Online Multiscale Dynamic Topic Models,“ KDD2010

購買行動解析のための
トピック追跡モデル

T. Iwata, S. Watanabe, T. Yamada, N. Ueda,
"Topic Tracking Model for Analyzing Consumer Purchase Behavior,“
IJCAI2009

Introduction

- 購買行動のモデリングは重要なタスク
  - 推薦システム
  - パーソナライズド広告
  - トレンド解析

Purchase behavior model

item　user　time

$$P(i \mid u,t)$$

文書：ユーザ
単語：商品

時刻tにユーザuが商品iを購入する確率

1982

Purchase behavior model

item　user　time

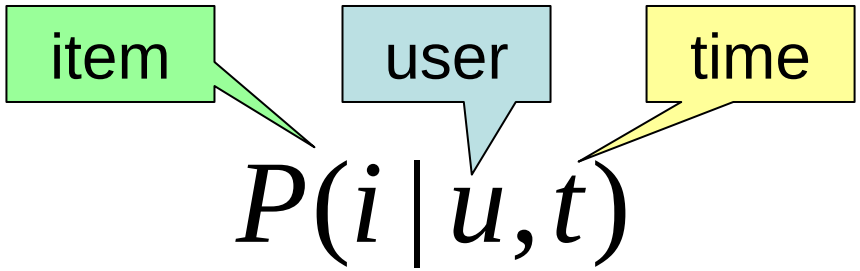
$$P(i \mid u, t)$$

時刻tにユーザuが商品iを購入する確率

1993

	A	B	C	…
Michael Jackson				
Beatles				
Deep purple				
Madonna				
⋮				

Purchase behavior model



時刻tにユーザuが商品iを購入する確率

2009

	A	B	C	…
Michael Jackson				
Beatles				
Deep purple				
Madonna				
⋮				

Purchase behavior model

item　user　time

$$P(i \mid u, t)$$

時刻tにユーザuが商品iを購入する確率

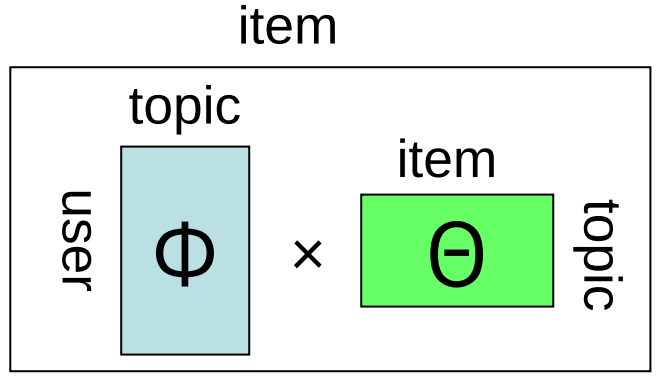
2009

	A	B	C	…
Michael Jackson				
Beatles				
Deep purple				
⋮				

パラメータ数＝ユーザ数×商品数×時刻数

Topic model

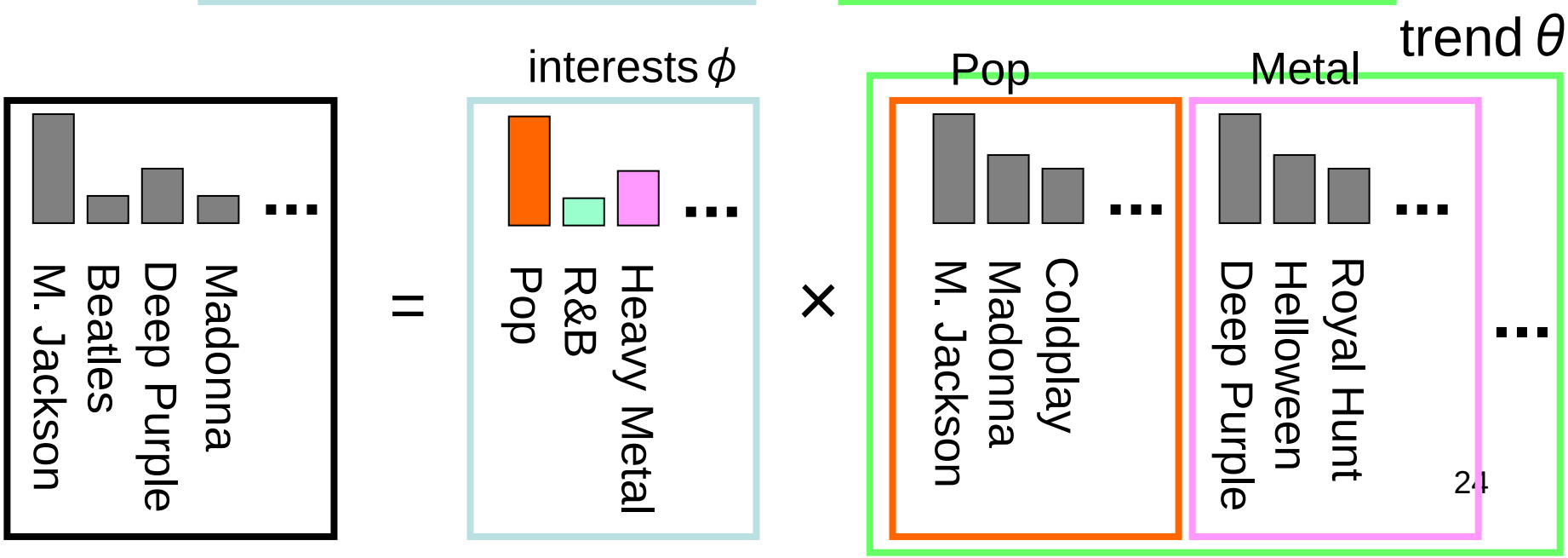
- Probabilistic latent semantic analysis
- パラメータ数を潜在トピックの導入により削減



$$P(i \mid u,t) = \sum_z P(z \mid u,t) P(i \mid z,t) = \sum_z \phi_{tzu} \theta_{tzi}$$

ユーザ毎の興味

トピック毎の流行



Bayesian extension

- Latent Dirichlet Allocation (LDA)
- 多項分布パラメータϕとθに対称ディリクレ事前分布を仮定
  - 推定がロバストに
  - 新たなユーザに確率を定義できる

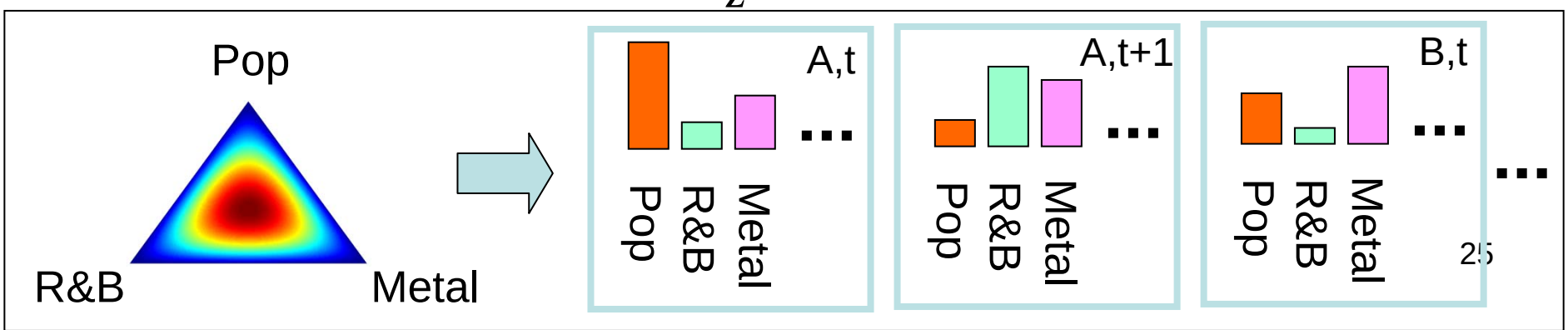
$$P(\phi_{tu}) \propto \prod_z \phi_{tuz}^{\gamma-1}$$

Interest prior

$$P(\theta_{tz}) \propto \prod_i \theta_{tzi}^{\gamma-1}$$

trend prior

$$P(i \mid u,t) = \sum_z \phi_{tzu}\theta_{tzi}$$



Bayesian extension

- Latent Dirichlet Allocation (LDA)
- 多項分布パラメータϕとθに対称ディリクレ事前分布を仮定
  - 推定がロバストに
  - 新たなユーザに確率を定義できる

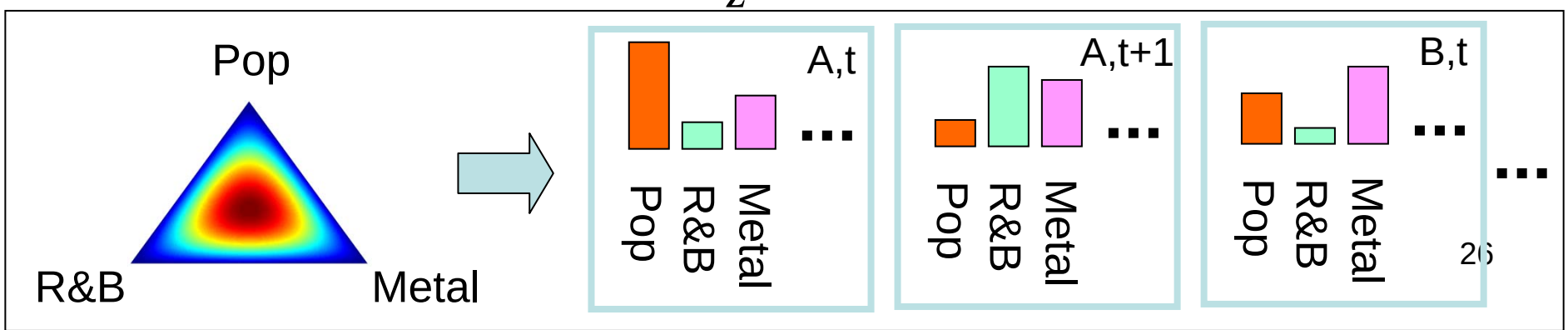
$$P(\phi_{tz}) \propto \prod \phi_{tzu}^{\gamma-1}$$

$$P(\theta_{tz}) \propto \prod \theta_{tzi}^{\gamma-1}$$

Inter … d prior

ダイナミクスが考慮されていない

$$P(i \mid u,t) = \sum_z \phi_{tzu} \theta_{tzi}$$



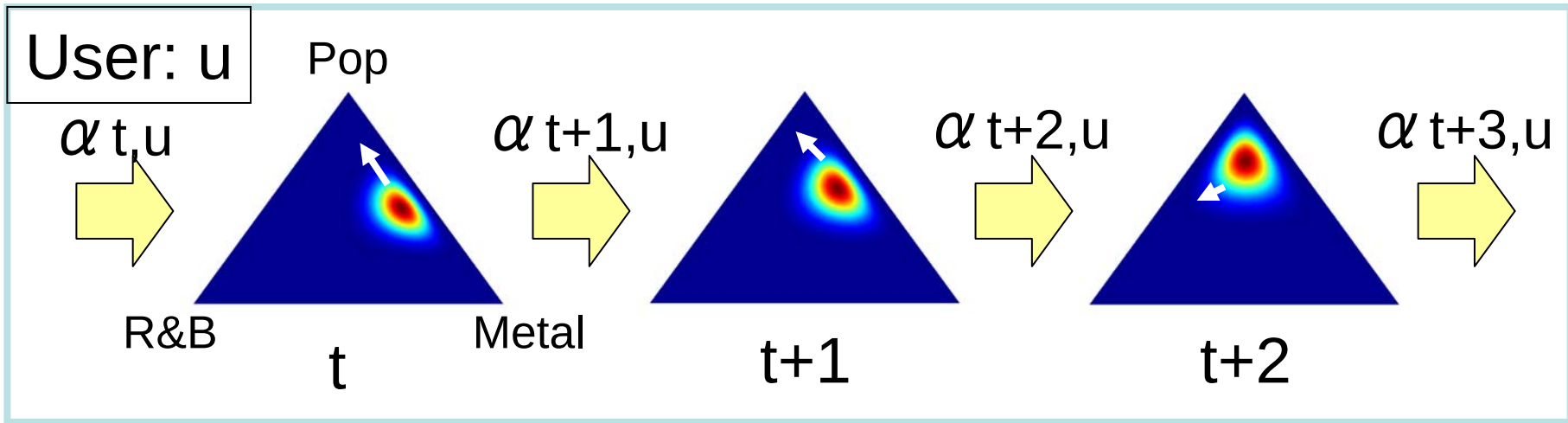
Topic Tracking Model (TTM)

- 購買行動を解析するためのトピックモデル
  - ユーザ興味とトピック流行にダイナミクスを導入

- TTMの特長
  - 変化への適応性
    - 興味や流行は時間的に変化する
  - 逐次学習
    - 膨大なログが日々蓄積される

Incorporate the dynamics of interests

- 興味の事前分布が過去の興味に依存
- 共役事前分布(Dirichlet) ⇒学習が簡単に



Interest prior $P(\phi_{t,u} \mid \hat{\phi}_{t-1,u}, \alpha_{t,u}) \propto \prod_z \phi_{t,u,z}^{\alpha_{t,u}\hat{\phi}_{t-1,u,z}-1}$

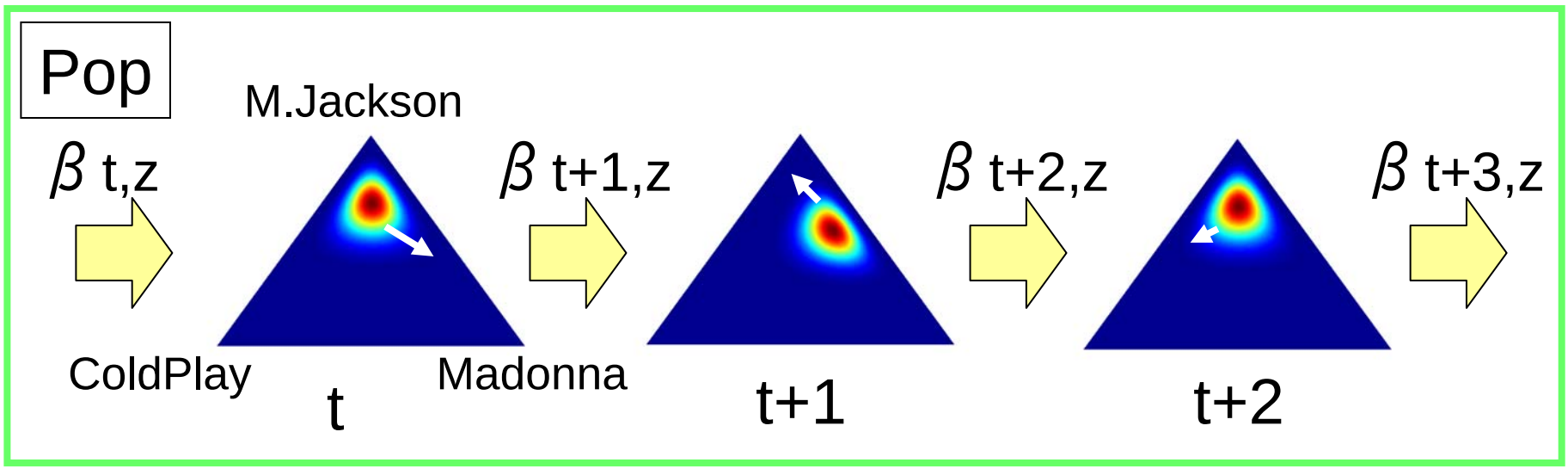
mean: previous interests $\phi_{t-1,u}$

precision: $\alpha_{t,u}$

興味の一貫性
（ユーザ興味がどれくらい変わりにくいか）

Incorporate the dynamics of trends

- 興味と同様に流行にもダイナミクスを導入



Trend prior $P(\theta_{t,z} \mid \hat{\theta}_{t-1,z}, \beta_{t,z}) \propto \prod_i \theta_{t,z,i}^{\beta_{t,z}\hat{\theta}_{t-1,z,i}-1}$

mean: previous trends θ t-1,z

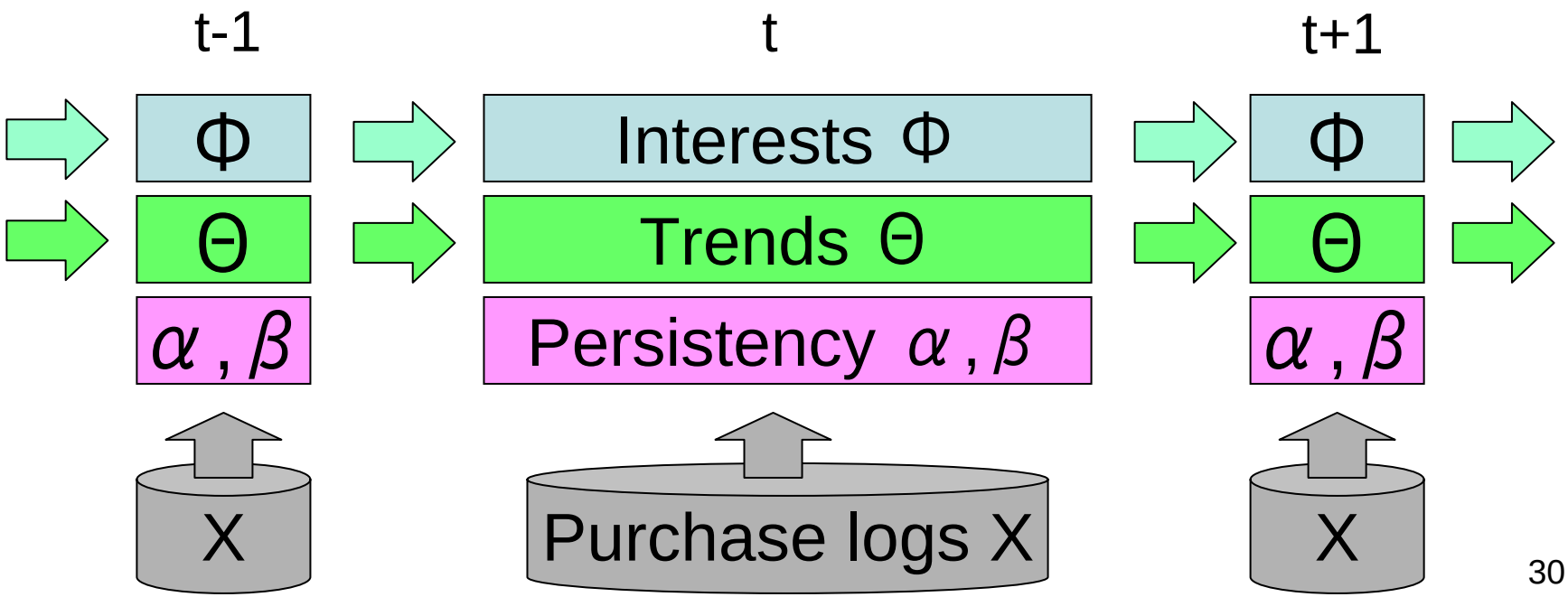
precision: β t,z

流行の一貫性
（トピック毎の流行がどれくらい変わりにくいか）

Online inference

- 確率的EMアルゴリズムを用いて逐次的に推定
  - 現在のデータのみを使用
    ⇒計算コストを削減
  - 過去のデータを保持する必要はない
    ⇒必要メモリを削減

a sample
(user, item, time)



Stochastic EM

E-step: collapsed Gibbs sampling of topic z

$$P(z_j = k|\boldsymbol{X}_t, \boldsymbol{Z}_{t\backslash j}, \hat{\boldsymbol{\Phi}}_{t-1}, \hat{\boldsymbol{\Theta}}_{t-1}, \boldsymbol{\alpha}_t, \boldsymbol{\beta}_t) \propto \frac{n_{t,u,k\backslash j} + \alpha_{t,u}\hat{\phi}_{t-1,u,k}}{n_{t,u\backslash j} + \alpha_{t,u}} \frac{n_{t,k,i\backslash j} + \beta_{t,k}\hat{\theta}_{t-1,k,i}}{n_{t,k\backslash j} + \beta_{t,k}}$$

M-step: maximum likelihood of α , β

$$\alpha_{t,u} \leftarrow \alpha_{t,u} \frac{\sum_z \hat{\phi}_{t-1,u,z}(\Psi(n_{t,u,z} + \alpha_{t,u}\hat{\phi}_{t-1,u,z}) - \Psi(\alpha_{t,u}\hat{\phi}_{t-1,u,z}))}{\Psi(n_{t,u} + \alpha_{t,u}) - \Psi(\alpha_{t,u})}$$

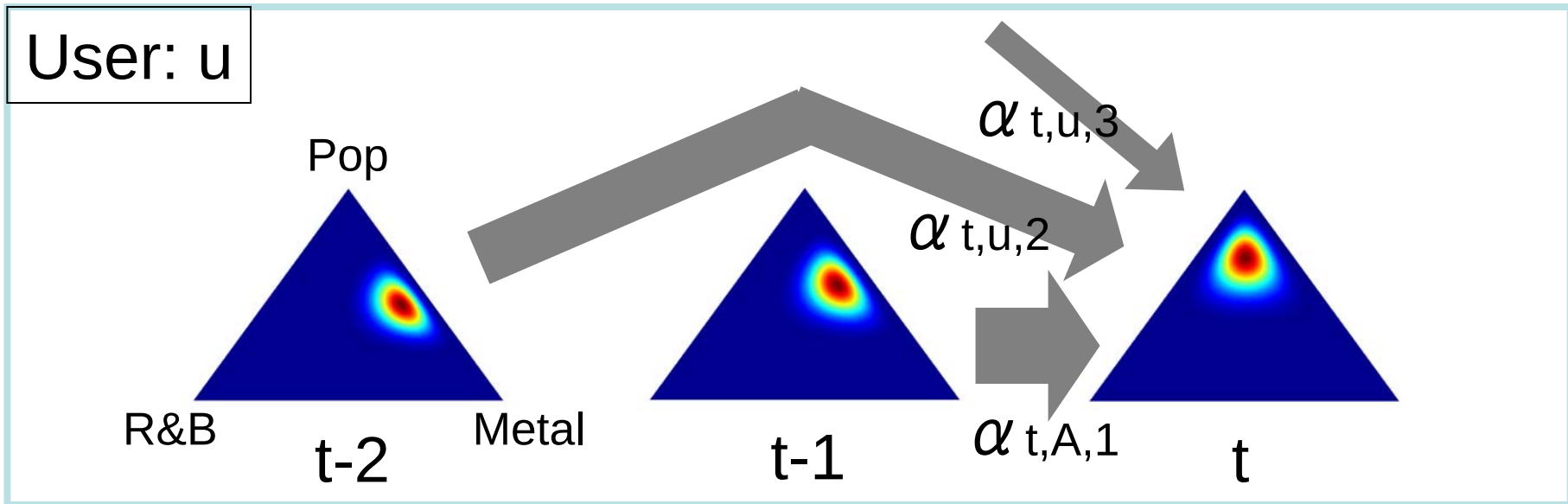
$$\beta_{t,z} \leftarrow \beta_{t,z} \frac{\sum_i \hat{\theta}_{t-1,z,i}(\Psi(n_{t,z,i} + \beta_{t,z}\hat{\theta}_{t-1,z,i}) - \Psi(\beta_{t,z}\hat{\theta}_{t-1,z,i}))}{\Psi(n_{t,z} + \beta_{t,z}) - \Psi(\beta_{t,z})}$$

Estimate ϕ and θ after the iteration of EM steps

$$\hat{\phi}_{t,u,z} = \frac{n_{t,u,z} + \alpha_{t,u}\hat{\phi}_{t-1,u,z}}{n_{t,u} + \alpha_{t,u}} \qquad \hat{\theta}_{t,z,i} = \frac{n_{t,z,i} + \beta_{t,z}\hat{\theta}_{t-1,z,i}}{n_{t,z} + \beta_{t,z}}$$

Capturing long term dependences

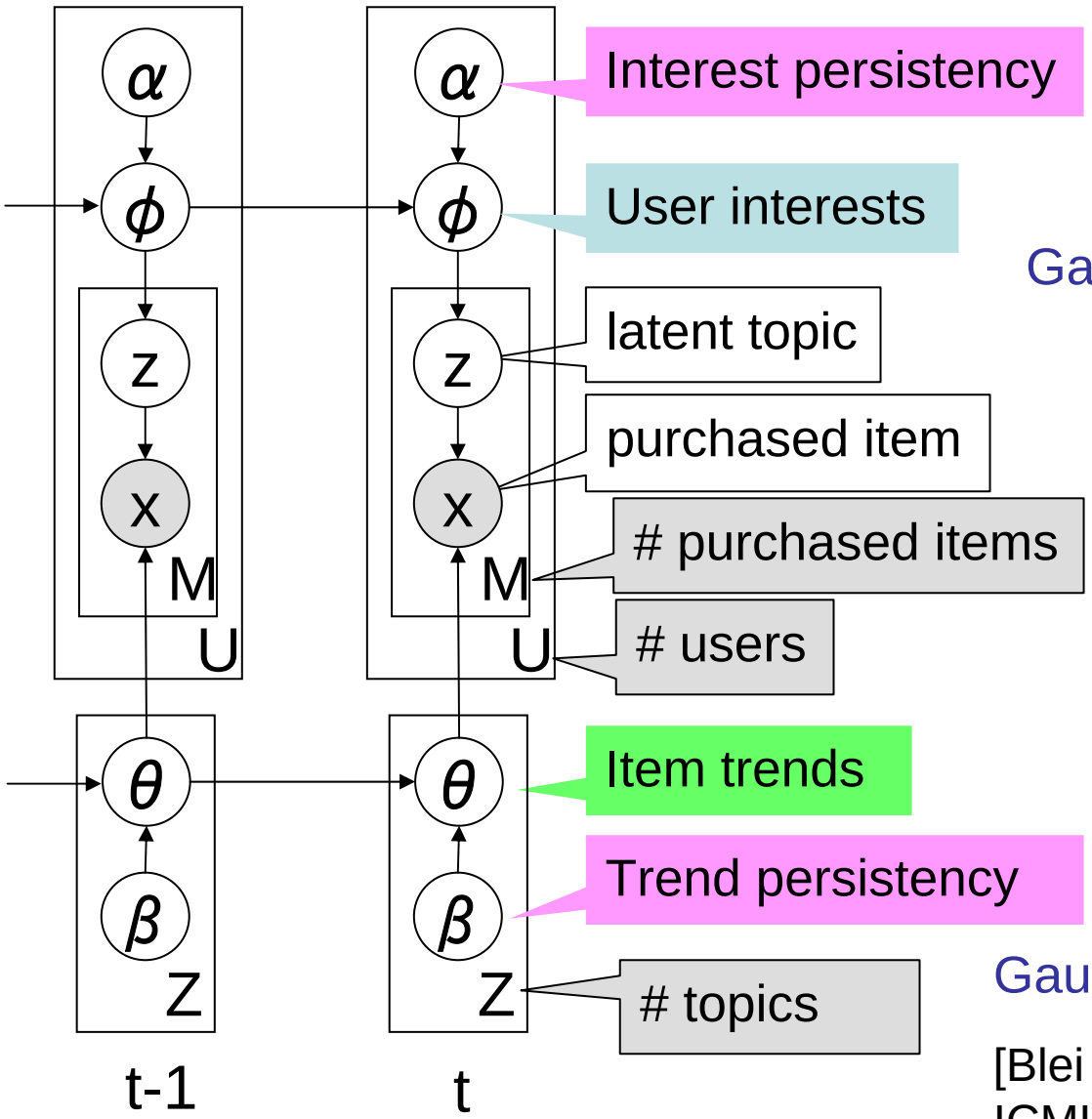
- 興味・流行を過去L期間の興味・流行に依存するように拡張



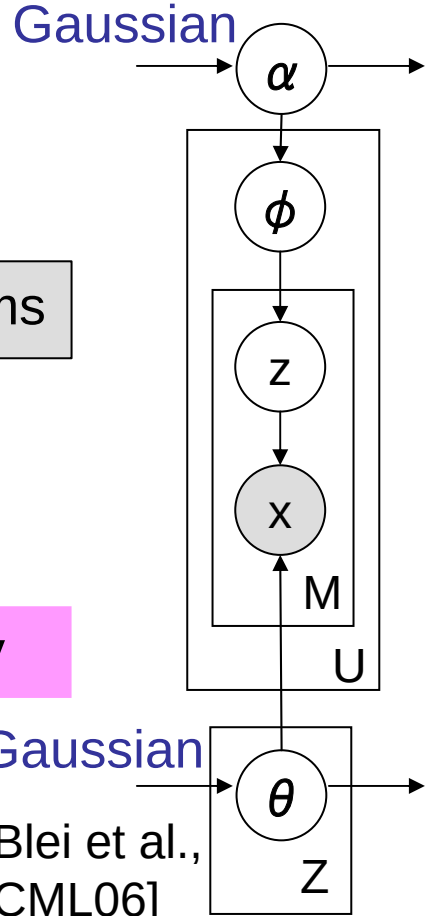
$$P(\phi_{t,u}|\{\hat{\phi}_{t-l,u},\alpha_{t,u,l}\}_{l=1}^{L}) \propto \prod_{z} \phi_{t,u,z}^{\sum_l \alpha_{t,u,l}\hat{\phi}_{t-l,u,z}-1}$$

- 長期依存性を考慮することで情報損失を削減
- 複数の推定値を利用するため学習がロバストに
- オンライン学習は同様に可能

Graphical models of TTM and related models

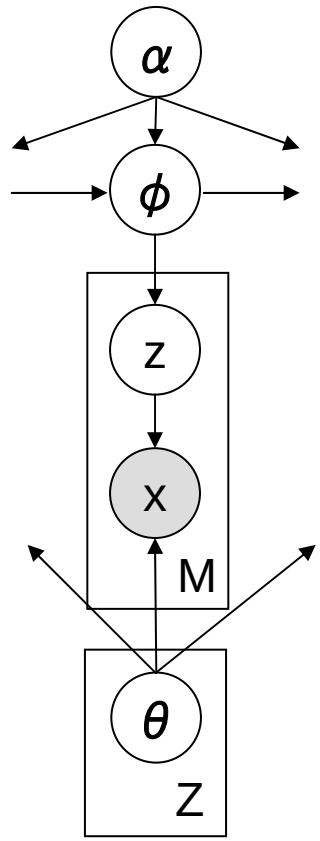


Dynamic Topic Model



[Blei et al., ICML06]

Dynamic Mixture Model



[Wei et al., IJCAI07]

Experiments

- 最後の購買を予測(それまでのデータを学習に利用)
- Movie
  - 110 days, 70,122 users, 7,469 items
  - 11,243,935 transactions
- Cartoon
  - 151 days, 143,212 users, 206 items
  - 12,642,505 transactions
- 比較手法
  - LDAall: 過去全てのデータを学習に使用
  - LDAonline: LDAのオンライン学習版
  - LDAone: 過去1日のデータのみを学習に使用

Average *N*-best accuracies (%)

(a) movie

N	LDAall	LDAonline	LDAone	TTM1	TTM10
1	1.21 (0.61)	1.08 (0.54)	1.91 (0.78)	2.22 (0.91)	**2.46** (0.92)
2	2.18 (0.79)	2.00 (0.78)	3.52 (1.22)	3.99 (1.33)	**4.47** (1.36)
3	3.06 (1.04)	2.81 (1.02)	5.04 (1.64)	5.60 (1.75)	**6.35** (1.85)
4	3.90 (1.27)	3.56 (1.24)	6.24 (1.90)	6.82 (2.01)	**7.82** (2.15)
5	4.70 (1.51)	4.26 (1.44)	7.37 (2.20)	7.92 (2.26)	**9.20** (2.42)

(b) cartoon

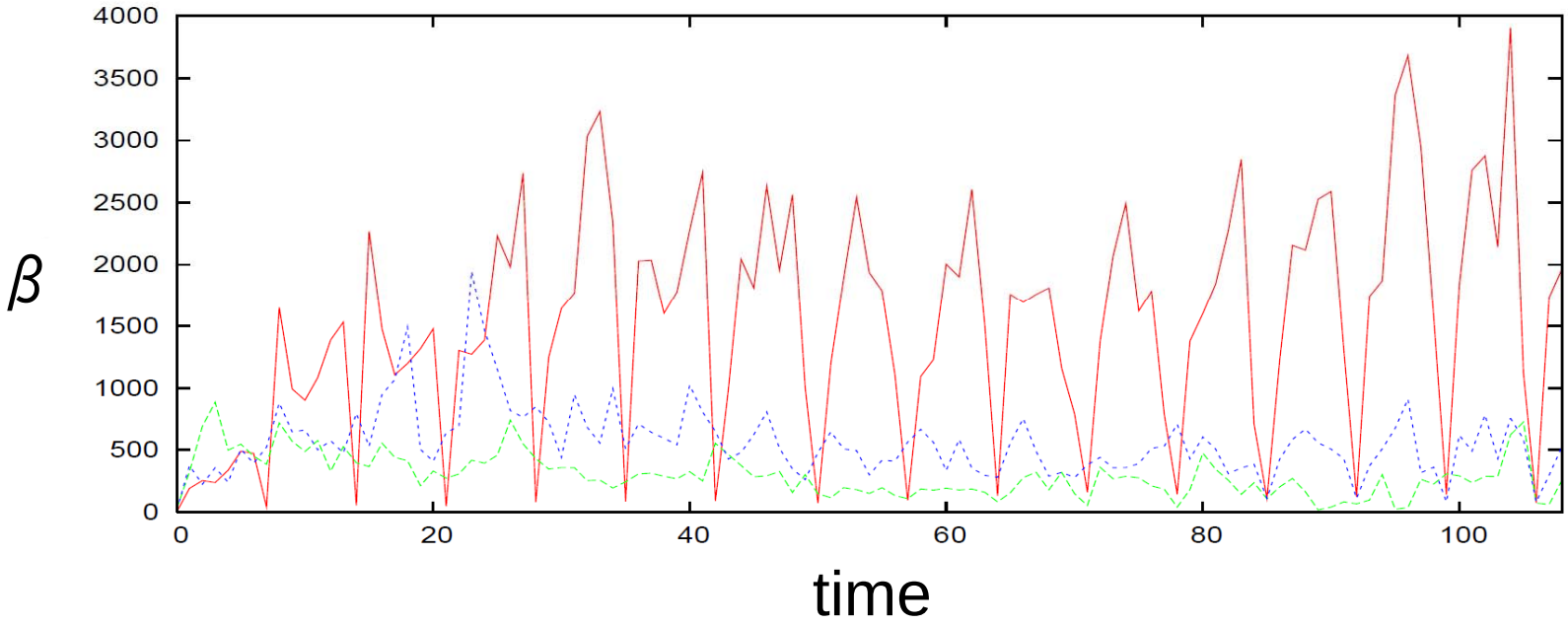
N	LDAall	LDAonline	LDAone	TTM1	TTM10
1	27.0 (3.3)	26.0 (3.5)	24.8 (4.5)	26.8 (4.2)	**30.5** (3.4)
2	37.3 (3.6)	35.1 (4.2)	32.4 (4.9)	34.2 (4.5)	**39.9** (3.5)
3	43.7 (3.9)	41.1 (4.8)	37.2 (5.3)	39.8 (4.6)	**45.9** (3.3)
4	48.5 (4.0)	45.8 (5.1)	40.9 (5.3)	44.5 (4.6)	**50.6** (3.2)
5	52.4 (4.2)	49.6 (5.4)	44.1 (5.4)	48.5 (4.6)	**54.4** (3.0)

Average computational time (sec)

	LDAall	LDAonline	LDAone	TTM1	TTM10
movie	12,380.3	523.7	503.2	455.3	707.0
cartoon	12,528.4	663.2	674.6	633.9	936.9

- LDAallは全てのデータを利用するため計算量大
- TTM10は1日あたり100,000購買からなるデータでも約12分で計算可能
- TTMでは依存距離に対し線形に計算量増加

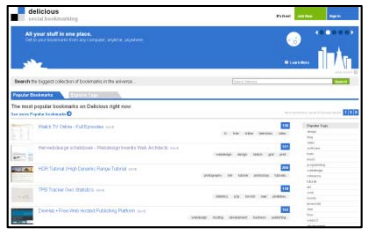
Examples of trend persistency β for each day



- 赤トピックは過去の流行に大きく依存
- 緑トピックは依存小
- 同じトピックでも依存性は時間的に変化する
  - 商品が新発売されると依存性が低くなる

Trend analysis in Delicious tags

- Deliciousデータ（2005/11から2008/3)
  - ソーシャルブックマークサービス
  - 購買商品→タグ
  - 1購買 が(user, tag, time)から成る
  - 37月, 15,466ユーザ, 14,586タグ



- 2008/11のトピック

politics,obama,election,election2008,mccain
mobile,iphone,google,microsoft,android
web2.0,twitter,socialmedia,tools,socialnetworking
mac,iphone,osx,software,apple

Topic “politics”

2007/8 politics,economics,history,news,culture
2007/9 politics,economics,history,news,war
2007/10 politics,economics,history,news,law
2007/11 politics,economics,history,media,news
2007/12 politics,economics,history,religion,government
2008/1 politics,economics,history,government,media
2008/2 politics,economics,history,obama,gov
2008/3 politics,economics,history,obama,war
2008/4 politics,economics,history,media,gove
2008/5 politics,economics,history,culture,gov
2008/6 politics,obama,economics,history,law
2008/7 politics,economics,obama,history,law
2008/8 politics,obama,economics,history,w
2008/9 politics,mccain,palin,election,obama
2008/10 politics,obama,election,mccain,econo
2008/11 politics,obama,election,election2008,mccain

June/3/2008 Obama was named the presumptive nominee.

August/29/2008 McCain announced that he had chosen Palin as his running mate.

Nov/4/2008 the U.S. Presidential Election

Topic “web2.0”

2007/1 web2.0,social,community,search,blogging,google,video,internet,secondlife,rss
2007/2 web2.0,community,social,video,rss,blogging,google,search,tools,yahoo
2007/3 web2.0,community,social,blogging,twitter,collaboration,google,search,rss,video
2007/4 web2.0,community,social,blogging,video,google,twitter,collaboration,rss,search
2007/5 web2.0,community,social,search,video,google,blogging,twitter,tools,socialnetwork
2007/6 web2.0,community,social,video,facebook,blogging,tools,google,socialnetworking,
2007/7 web2.0,facebook,community,socialnetworking,social,blogging,tools,video,google,
2007/8 web2.0,social,socialnetworking,facebook,community,video,blogging,google,tools,
2007/9 web2.0,facebook,socialnetworking,social,google,community,tools,blogging,video,
2007/10 web2.0,facebook,socialnetworking,social,google,community,blogging,video,tools
2007/11 web2.0,facebook,socialnetworking,social,google,community,advertising,opensoc
2007/12 web2.0,facebook,socialnetworking,social,google,blogging,twitter,community,vide
2008/1 web2.0,socialnetworking,facebook,twitter,social,socialmedia,blogging,tools,comm
2008/2 web2.0,socialnetworking,twitter,social,community,google,socialmedia,facebook,to
2008/3 web2.0,twitter,socialnetworking,social,socialmedia,community,blogging,facebook,
2008/4 web2.0,twitter,socialnetworking,social,socialmedia,blogging,tools,community,face
2008/5 web2.0,twitter,socialnetworking,social,socialmedia,community,tools,blogging,face
2008/6 web2.0,twitter,socialnetworking,socialmedia,social,tools,blogging,community,face
2008/7 web2.0,twitter,socialmedia,socialnetworking,social,tools,blogging,community,goog
2008/8 web2.0,twitter,tools,socialmedia,socialnetworking,social,blogging,search,faceboo
2008/9 web2.0,twitter,socialmedia,socialnetworking,tools,social,google,blogging,search,t
2008/10 web2.0,twitter,socialmedia,tools,socialnetworking,social,blogging,marketing,goo
2008/11 web2.0,twitter,socialmedia,tools,socialnetworking,social,facebook,microblogging

トピック追跡モデルまとめ

- 購買履歴解析のためのトピックモデルを提案
  - 高い予測精度
  - 高い計算効率
  - トレンド解析可能
- 今後の課題
  - トピック数の自動推定（新トピック検出）
    - Teh et al., Hierarchical Dirichlet processes, Journal of the American Statistical Association, 2006
    - Morinaga and Yamanishi, Tracking Dynamics of Topic Trends using a Finite Mixture Model, KDD2004
  - 依存時間の自動推定
    - Wang et al., Continuous Time Dynamic Topic Models, UAI2008
    - Iwata et al., Online Multiscale Dynamic Topic Models, KDD2010

最後に

- トピックモデル解説
  - 幅広い応用範囲
  - 拡張が容易
  - 実装が簡単
- トピックモデル応用
  - 購買行動解析