Wikipedia のための Latent Dirichlet Allocation

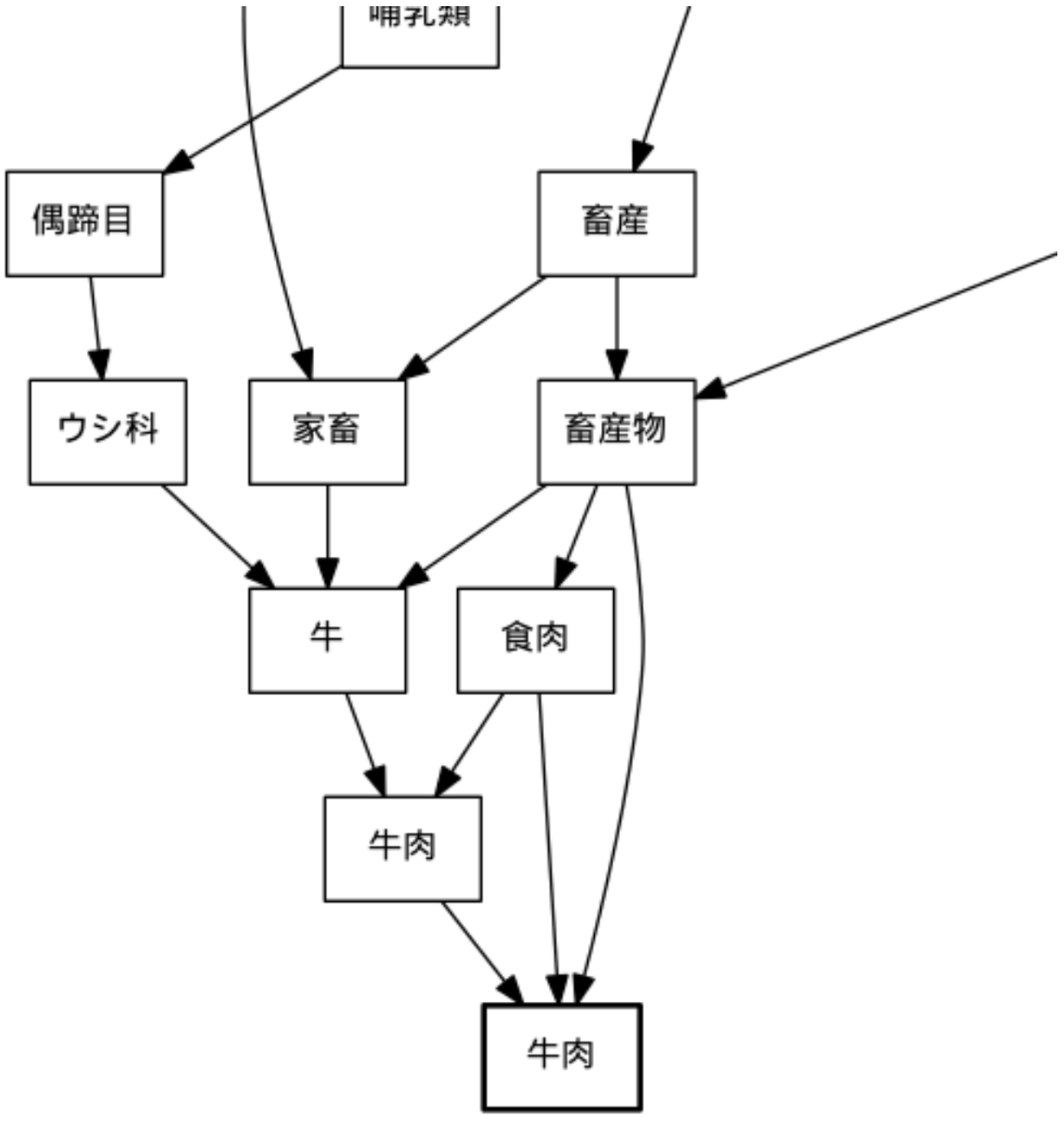
小松 弘佳

モチベーション

- Wikipedia のカテゴリー情報を使ってトピックをより厳密に取りたい
  - 教師ありトピックモデルが必要
- 好きな粒度のトピックがほしい
  - 宇宙>惑星>木星
  - 階層的なトピック
- **教師ありの階層的トピックモデル**

モチベーション

- 文章がどんな Wikipedia カテゴリーを持つか調べれば深い意味情報が分かる
- 質問応答の Question Classification などでの応用が目的

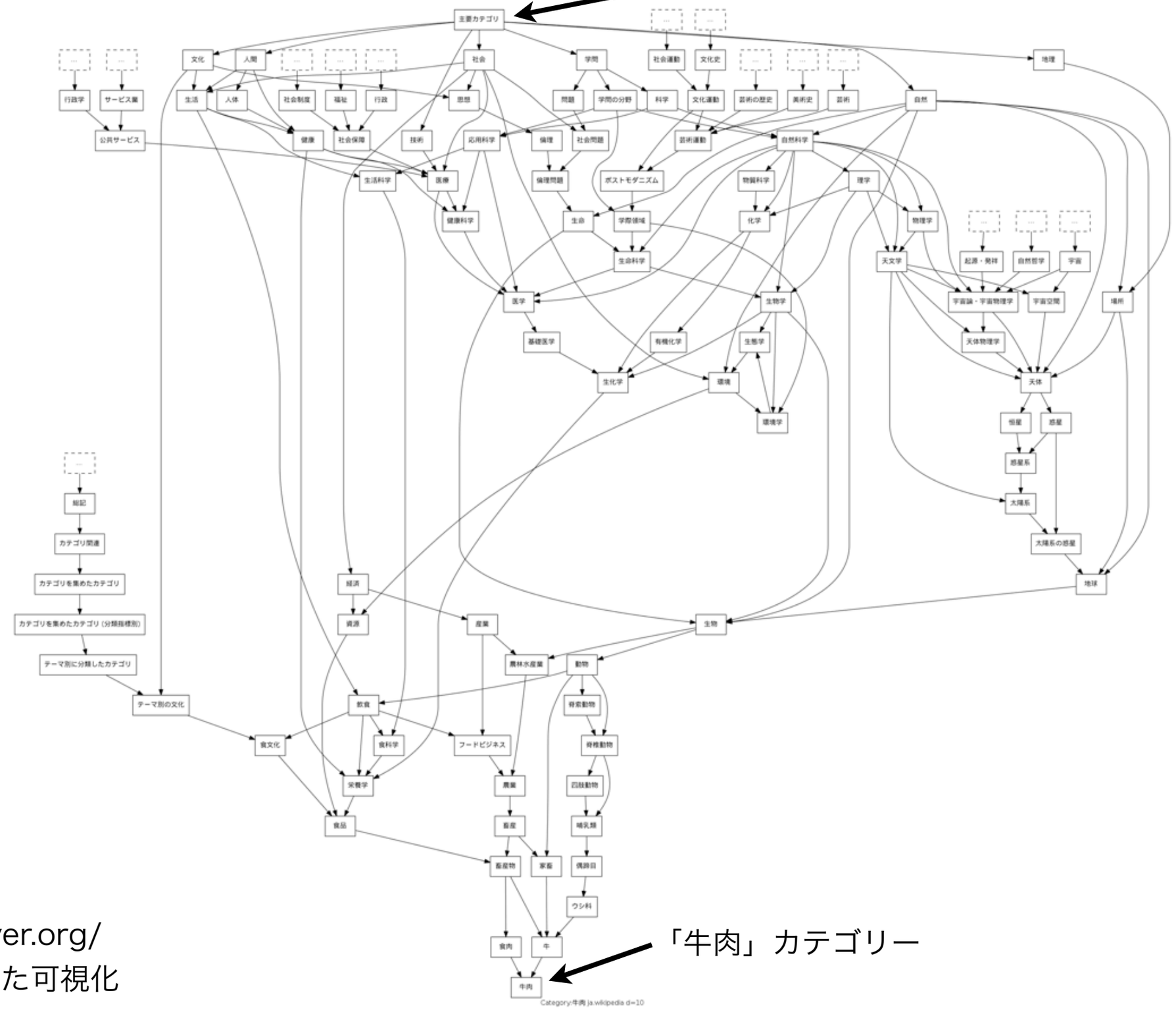


Wikipedia のデータの特徴

- 複雑なカテゴリー体系
- 100個のルートカテゴリーからそれぞれの記事に複数のルートで行ける
- ただしループに落ち入ることはない
- 記事はどの葉ノードからも生成される
  （記事のカテゴリーは最後の層まで必ず行くわけではない）
- どのルートを通るかによってノードの数が可変長
- 基準はある程度明確
  - http://ja.wikipedia.org/wiki/Wikipedia:カテゴリの方針

Wikipedia のデータの特徴

「牛肉」カテゴリにたどり着くまでのルート（一部）

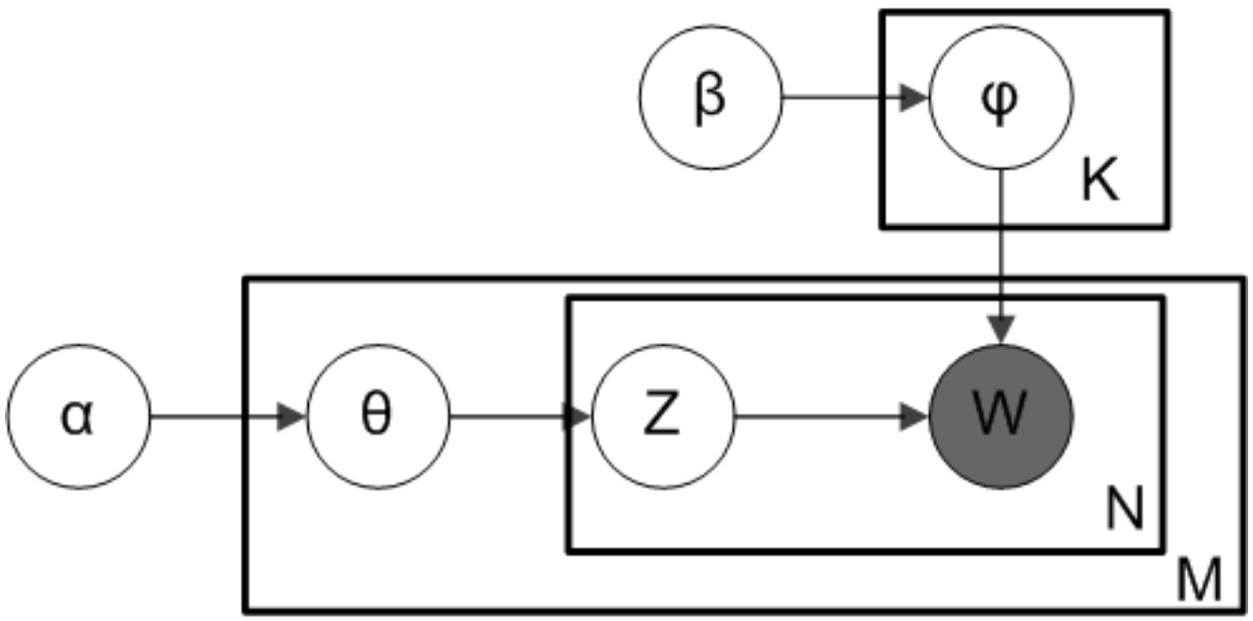


「牛肉」カテゴリー

Catgraph (http://toolserver.org/~dapete/catgraph/) を用いた可視化

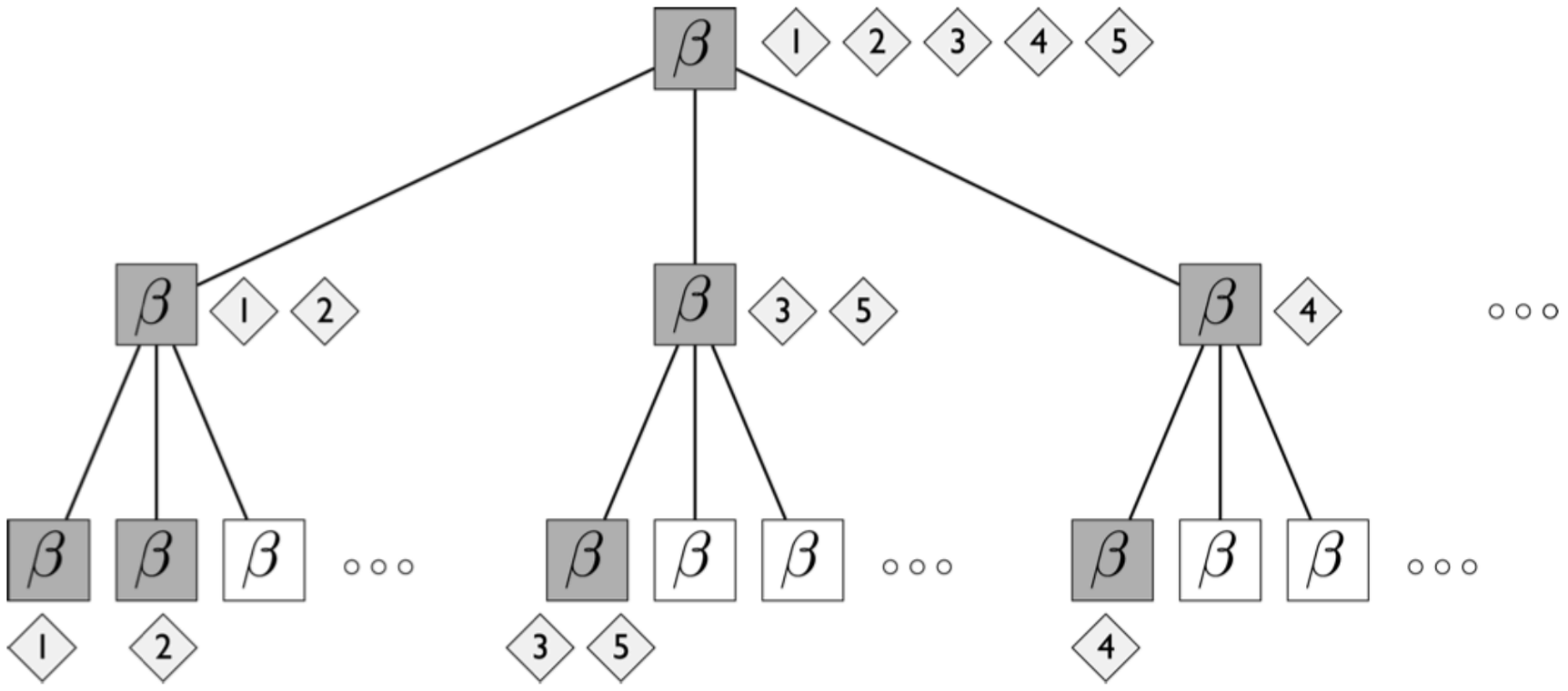
Latent Dirichlet Allocation (Blei+ 03)

- 教師なしのトピックモデル
  - それぞれのワードにトピックを付ける
  - 文章にもトピックをひも付けることが可能



関連する LDA の亜種 (1)

- 階層化された LDA
  - Hierarchical LDA (Blei+ 2010) (Blei+ 2003)
    - 教師なしで1層からL層までのトピックを構築・学習する
    - 任意の層からトピックを推定できる（好きな粒度のトピックが取れる）



Hierarchical LDA

- nCRP (nested Chinese Restaurant Process) が使われている
  - CRP の階層化
- どの階層から単語が生成されたかを表す行列 $z_{dn} \in \{0, ..., L\}$ がある
- 単語は 0 から z_{dn} までそれぞれの階層でトピックが割り当てられている

(1) For each table $k \in \mathcal{T}$ in the infinite tree,
  (a) Draw a topic $\beta_k \sim$ Dirichlet(η).

(2) For each document, $d \in \{1, 2, \ldots, D\}$
  (a) Draw $\mathbf{c}_d \sim$ nCRP(γ).
  (b) Draw a distribution over levels in the tree, $\theta_d \mid \{m, \pi\} \sim$ GEM(m, π).
  (c) For each word,
    i. Choose level $Z_{d,n} \mid \theta_d \sim$ Discrete(θ_d).
    ii. Choose word $W_{d,n} \mid \{z_{d,n}, \mathbf{c}_d, \beta\} \sim$ Discrete($\beta_{\mathbf{c}_d}[z_{d,n}]$), which is parameterized by the topic in position $z_{d,n}$ on the path $\mathbf{c}_d$.

関連する LDA の亜種 (2)

- 教師あり LDA
  - カテゴリー情報を教師ありデータとして与える手法
  - これ以外にもたくさんの手法が考案されている
- Labeled LDA(Ramage+ 2009)
  - トピックとカテゴリ情報が一対一で対応
  - 今回はこれを拡張
- Dirichlet Process with Mixed Random Measures (Kim+ 2012)
  - トピックとカテゴリが複数対応

Labeled LDA

- トピックはそれぞれ1つのラベルに対応しているモデル
- ハイパーパラメータ α は、ラベルの観測列 Λ に依存している

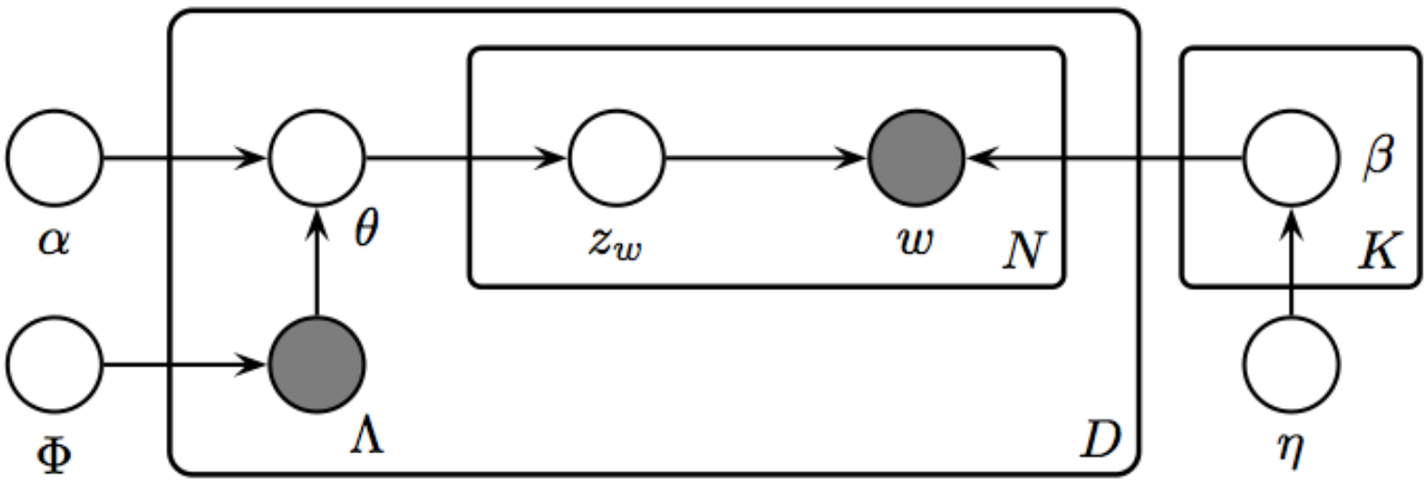


Figure 1: Graphical model of Labeled LDA: unlike standard LDA, both the label set $\boldsymbol{\Lambda}$ as well as the topic prior $\boldsymbol{\alpha}$ influence the topic mixture $\boldsymbol{\theta}$.

Labeled LDA

- 教師ありデータとして文書に与えるカテゴリー情報は、複数のラベルでもOK
- 実装が簡単（LDAにΛの部分の処理を加えるだけ）

$$\beta_k \sim Dir(\eta)$$
$$\Lambda_k^{(d)} \sim Bernouli(\Phi_k)$$
$$\theta^{(d)} \sim Dir(\alpha^{(d)})$$
$$\text{where} \quad \alpha^{(d)} = (\alpha_k)_{\left\{k|\Lambda_k^{(d)}=1\right\}}$$
$$\mathrm{z}_i^{(d)} \sim Multi(\theta^{(d)})$$
$$\mathrm{w}_i^{(d)} \sim Multi(\beta_{z_i^{(d)}})$$

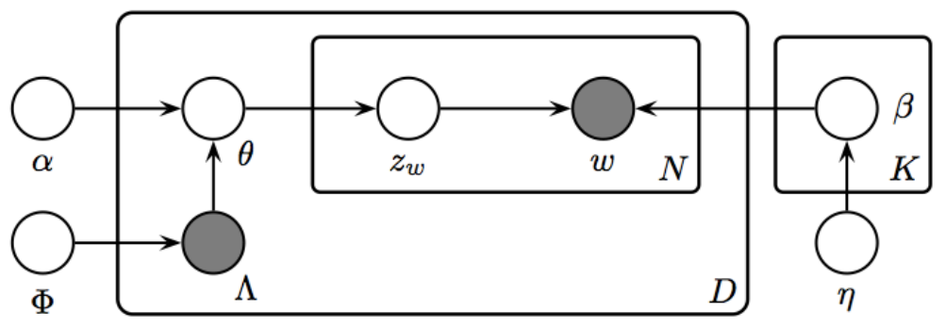


Figure 1: Graphical model of Labeled LDA: unlike standard LDA, both the label set $\mathbf{\Lambda}$ as well as the topic prior $\boldsymbol{\alpha}$ influence the topic mixture $\boldsymbol{\theta}$.

Labeled LDA

- 教師ありデータとして文書に与えるカテゴリー情報は、複数のラベルでもOK
- ラベルが違うものは i.i.d であるという仮定
- Labeled LDA は1文書に複数のカテゴリーを全くひも付けないと、ただのナイーブベイズになる
- ラベルは1つの場合も多いが…
  - カテゴリーであれば、Wikipedia のような複雑なデータほどルートが複数あり性能が発揮できそう

提案手法

- Wikipedia のような複雑なカテゴリー体系でも使える、教師ありの階層的な LDA
- カテゴリの行列表現

$$\mathrm{E}_{i,\ell,k} \in \{0, 1\}$$

  - i文書のl階層のkカテゴリーが出現したら1、しなかったら0
  - これを Hierarchical LDA の式に放り込む

提案手法

- hLDA の nested CRP の式を、新たに given となるカテゴリー情報 E を加えて改変

hLDA

$$\mathrm{P}\left(\mathrm{r}_{i\ell} = x \middle| r_{-i}, r_{i,1:(\ell-1)}\right) = \begin{cases} \dfrac{|j \neq i | r_{j,1:(\ell-1)} = r_{i,1:(\ell-1)}, r_{j\ell} = x|}{|j \neq i | r_{j,1:(\ell-1)} = r_{i,1:(\ell-1)}| + \gamma_\ell} & \text{if x is an existing branch,} \\ \dfrac{\gamma_\ell}{|j \neq i | r_{j,1:(\ell-1)} = r_{i,1:(\ell-1)}| + \gamma_\ell} & \text{if x is a new branch} \end{cases}$$

提案手法

$$P\left(r_{i\ell} = x \middle| r_{-i}, r_{i,1:(\ell-1)}), E\right) = \begin{cases} \dfrac{|j \neq i | r_{j,1:(\ell-1)} = r_{i,1:(\ell-1)}, E_{j,1:(\ell-1)} = E_{i,1:(\ell-1)}, r_{j\ell} = x|}{|j \neq i | r_{j,1:(\ell-1)} = r_{i,1:(\ell-1)}, E_{j,1:(\ell-1)} = E_{i,1:(\ell-1)}| + \gamma_\ell} & \text{if existing,} \\ \dfrac{\gamma_\ell}{|j \neq i | r_{j,1:(\ell-1)} = r_{i,1:(\ell-1)}, E_{j,1:(\ell-1)} = E_{i,1:(\ell-1)}| + \gamma_\ell} & \text{if new} \end{cases}$$

提案手法

- 同じ階層に複数のノードがある

  ✓ nCRP によって一番確率の高いルート1つに絞られる

- 記事はどのノードからも生成される

  ✓ カテゴリーがない階層は $E_{i,\ell}$ の中身を全て0にする

- どのルートを通るかによってノードが可変長

  ✓ L を一番遠回りしたとき通るノードの数にする

Future works

- 実験
- 提案手法は、今まで扱えなかったような複雑だが意味情報を含んだ階層カテゴリーの分類ができるようになった
    - 複雑なカテゴリー情報は実際に効くのか NLP の何らかのタスクのシステムに組み込んだ形で試す必要がある