2015 年度修士論文要旨

# 相関トピックモデルによる文書分類についての一考察

関西学院大学大学院理工学研究科
数理科学専攻　森本研究室　中谷陽平

大量のデータを活用し、有益な情報を発見するためのツールとして注目されているのがトピックモデルである．トピックモデルは文書データの解析手法として提案された．トピックモデルを用いることにより，人出を介在させることなく，大量の文書集合から話題になっているトピックを抽出したり，それぞれの文書がどのようなトピックを持っているかがわかる．更にトピックに相関があると考えられる場合がある．例えば新聞記事の場合，政治と経済の 2 つのトピックを持つ記事は多くなるが，政治と芸能の 2 つのトピックを持つ記事は少なくなる．それを仮定したモデルが Li and McCallum(2006) の相関トピックモデルであり、それらを用いて文書分類を行っていく．

## 1　Latent Dirichlet Allocation(LDA)

代表的なトピックモデルである LDA では，1 つの文書が複数のトピックを持つと仮定する．具体的な文書集合の生成過程は以下の通りである．

LDA の生成過程

1. For トピック $k = 1, \cdots, K$
   （a）単語分布を生成 $\boldsymbol{\phi}_k \sim \mathrm{Dirichlet}(\beta)$
2. For 文書 $d = 1, \cdots, D$
   （a）トピック分布を生成 $\boldsymbol{\theta}_d \sim \mathrm{Dirichlet}(\alpha)$
   （b）For 単語 $n = 1, \cdots, N_d$
      i. トピックを生成 $z_{dn} \sim \mathrm{Categorical}(\boldsymbol{\theta}_d)$
      ii. 単語を生成 $w_{dn} \sim \mathrm{Categorical}(\boldsymbol{\phi}_{z_{dn}})$

## 2　相関トピックモデル

相関トピックモデルは，トピックに階層構造を導入することによって，トピック間の関係をモデル化する．たとえば，「料理」「健康」「保険」「薬」の 4 つの下位トピックがあるとする．「料理」「健康」トピックは同じ文書で出てくることが多く，また「健康」「保

険」「薬」はよく一緒に議論されるトピックである．上位トピックとして「料理・健康」，「健康・保険・薬」を用意することにより，これらの下位トピックの関係性をモデル化できる．具体的には，単語ごとにまず上位トピック分布 $\boldsymbol{\theta}_d$ を用いて，上位トピック $y_{dn}$ を選ぶ．次に，その上位トピックに応じた下位トピック分布 $\boldsymbol{\theta}_{d,y_{dn}}$ を用いて下位トピックを選ぶ．そして，選んだ下位トピックの単語分布 $\boldsymbol{\phi}_{z_{dn}}$ に従って語彙が決められる．文書 $d$ の上位トピック分布は $\boldsymbol{\theta}_d = (\theta_{d1}, \cdots, \theta_{dS})$ であり，$\theta_{ds}$ は文書 $d$ で上位トピック $s$ が選ばれる確率，$S$ は上位トピック数を表す．また，下位トピック分布は文書ごとに $S$ 個あり，$\boldsymbol{\Theta}_d = (\boldsymbol{\theta}_{d1}, \cdots, \boldsymbol{\theta}_{dS})$ と表す．上位トピックに応じた下位トピック分布を用いることにより，同じ文書に現れやすい下位トピックをモデル化できる．生成過程は以下の通りである．

相関トピックモデルの生成過程

1. For トピック $k = 1, \cdots, K$
   （a）単語分布を生成 $\boldsymbol{\phi}_k \sim \text{Dirichlet}(\beta)$
2. For 文書 $d = 1, \cdots, D$
   （a）上位トピック分布を生成 $\boldsymbol{\theta}_d \sim \text{Dirichlet}(\alpha_0)$
   （b）For $s = 1, \cdots, S$
      i. 下位トピック分布を生成 $\theta_{ds} \sim \text{Dirichlet}(\alpha_s)$
   （c）For 単語 $n = 1, \cdots, N_d$
      i. 上位トピック分布を生成 $y_{dn} \sim \text{Categorical}(\boldsymbol{\theta}_d)$
      ii. 下位トピック分布を生成 $y_{dn} \sim \text{Categorical}(\boldsymbol{\theta}_{d,y_{dn}})$
      iii. 単語を生成 $w_{dn} \sim \text{Categorical}(\boldsymbol{\phi}_{z_{dn}})$

## 参考文献


[1] W.Li and A.McCallum， Pachinko allocation： Dag-structured mix-true models of topic correlations，In Proceedings of International Conference on Machine learning，ICML，2006，pp.577-584

[2] Blei，D.M.，Ng，A.Y.and Jordan，M.I.：Latent Dirichlet Allocation，Journal of Machine Learning Research，3，2003，pp.993-1022

[3] 岩田具治，トピックモデル，講談社，2015