1F3-5

中華レストラン過程へのディリクレ森分布による制約知識の導入

Introducting Constraint Knowledge based on Dirichlet Forest Distribution to Chinese Restaurant Process

立川華代　　小林一郎
Kayo Tatsukawa　　Ichiro Kobayashi

お茶の水女子大学大学院 人間文化創成科学研究科 理学専攻 情報科学コース
Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University

We have recently had many chances to treat a huge amount of documents. A lot of studies about extracting latent topics in documents have been done, which is accomplished by Latent Dirichlet Allocation(LDA). That is not done by using superficial information. When we extract topics by LDA, it sometimes happens that the words we assume that they should be in the same topic are devided into different topics. Therefore Andrzejewski and Hu have proposed methods in that we select some words which should be in the same topic and incorporate the words into LDA process as constrained knowledge. But the knowledge is constructed from subjective view of users in many cases and is not constructed from target documents automatically. We therefore construct constrained knowledge from documents and then consider how the constrained knowledge is applied to Chinese Restaurant Process.

1. はじめに

近年，大量のテキストの処理において文書の潜在的意味を推定する潜在的ディリクレ配分法 (LDA:Latent Dirichlet Allocation[2]) が使われ，多くの研究がされている．しかし，LDA では文書内に潜在しているトピック数を予め与えることによって潜在トピックを推定する必要がある．実際は文書の潜在トピック数は未知のため，与えるトピック数が推定結果に大きな影響を与えると考えられる．そこで，Teh ら [4] によって，階層ディリクレ過程を用いたディリクレ配分法 (HDP-LDA) が考案されている．ここでは階層ディリクレ過程を用いて，トピック数 K を対象に表現した K 項のディリクレ多項分布モデルにおいて，$K \to \infty$ としたモデルとして中華レストラン過程 (CRP:Chinese Restaurant Process) が提案された．

一方，本来同じトピックに高い確率で割り当てられるべきである単語同士が異なるトピックに割り当てられてしまうことが多々あり，これに対して Andrejewski ら [1] や Hu ら [3] では，通常の LDA で使われているディリクレ分布をディリクレ森分布に変えることで，単語の出現に関して制約知識を与える潜在トピック推定手法を提案している．本研究では，トピック数を推定するための中華レストラン過程を階層的にすることで，トピック数を事前に与える必要がなく，単語同士の制約知識を踏まえた潜在トピック推定法を提案する．

2. 中華レストラン過程

中華レストラン過程とはディリクレ過程 (DP:Dirichlet Process) を用いたクラスタ推定法の一つである．語彙を客，テーブルをクラスタと見立てて，テーブルの数がトピック数となる．レストランには無限個のテーブルがあると仮定し語彙である客が一人ずつ自分が座るテーブルを決定する．この際，客は既に他の客が座っているテーブルに座るかもしくは誰も座っていない未使用の新しいテーブルに座ることになる．既に他の客が座っているテーブルを選択する確率は，そのテーブルに座っている客の人数に比例することが知られている．これを式 (1) に示す．

$$p(z_i = k | z_1, ..., z_{i-1}) = \begin{cases} \dfrac{m_k}{\gamma + i - 1} & (k = 1, \cdots K) \\ \dfrac{\gamma}{\gamma + i - 1} & (k = K + 1) \end{cases} \quad (1)$$

ここで K は現在までのクラスタ数，k は推定されるトピックであり，m_k は z_1 から z_{i-1} の中でその値が k に等しいものの個数，つまり $i-1$ 番目までの語で既にそのテーブルに座っている客の人数である．$k = 1, ..., K$ の時，このトピックが選ばれる確率は m_k に比例する．また，$\gamma(\gamma > 0)$ はディリクレ分布のハイパーパラメータであり，この γ に比例する確率でまだ誰も座っていない新しいテーブルに座ることになる．最終的に，一人以上の客が座っているテーブルの個数がトピック数として推定される．

3. 制約付き中華レストラン過程

本研究では制約知識を導入するために通常の中華レストランに制約知識用の特別個室を用意する．制約知識を「同じトピックに入って欲しい単語群」とし，二層のディリクレ森分布を導入し，ディリクレ森分布では図 1 に示されるように通常の LDA によって確率割り当てがされている単語の中に制約が含まれる．制約が選択されると，さらにそこから新しいディリクレ分布で別室のテーブルに確率割り当てを行う．この時，一層目において制約によって通常のテーブルの個数が減る分をハイパーパラメータに個数の重みをつけることで下階層までのテーブルの出現確率を層間において影響がないように補填している．また，制約下のハイパーパラメータは「特別室」に置かれた関連性の強いクラスタ同士が高い確率で選択されるようにハイパーパラメータ η の値を高くする．

$$P(\pi, \phi \mid \gamma, \eta, K, C) = Dir(\pi; (\frac{\gamma}{K})_{k \notin C}, \frac{C_k \gamma}{K})_{k \in C}) \times \prod_{k \in C} Dir(\phi_k, \eta) \quad (2)$$

連絡先: 立川華代，お茶の水女子大学大学院人間文化創成科学研究科理学専攻情報科学コース，〒 112-8610 東京都文京区大塚 2-1-1，03-5978-5708，tatsukawa.kayo@is.ocha.ac.jp

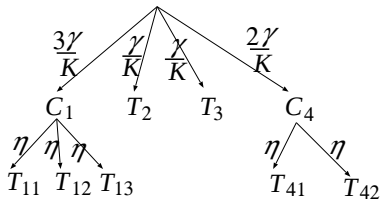

図 1: ディリクレ森分布

二層のディリクレ分布の式は式 (2) の確率式で表される．右辺第一項は一層目，第二項が二層目の確率を表現している．通常の中華レストラン過程によるトピックの推定式に制約 C と制約内でのテーブル番号 j を導入すると

$$P(z_i = (k,j) \mid z_1^{i-1}, \gamma, \eta, K, C) = \int P(z_i = (k,j) \mid \pi, \phi) p(\pi, \phi \mid z_1^{i-1}, \gamma, \eta, K, C) d\pi d\phi \quad (3)$$

となる．

ここで制約知識用の個室付きレストランの説明をする．レストラン内には制約単語ではない通常の単語のお客が座る大広間のテーブルと，制約単語のお客が入る個室とテーブルを用意する．図 2 の上方の三つの部屋が個室を表現している．レストランに含まれる全テーブル数を K，個室に含まれるテーブルの総数を k_c，個室ごとのテーブル数を $C_k (K_c = \sum_{k \in C} C_k; k$ は個室番号)，制約の集合を C とする．図 2 では $K = 15$, $K_c = 7$, $C_1 = 2, C_2 = 3, C_3 = 2, C = \{1,2,3\}$ となる．単語である客は，大広間に入るか，個室に入るかを選択し，大広間の場合は大広間のテーブルに座る．個室の場合は，その選ばれた個室内のテーブルに座る．

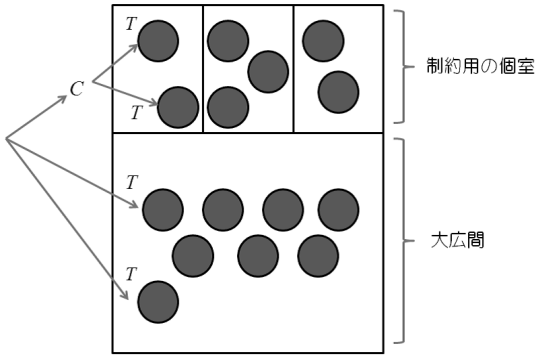

図 2: レストランの室内

CRP では式 (1) のようにテーブルに座っている人数に比例して既にあるテーブルに座るか，パラメータの値に比例して新しいテーブルに座る．そこでこの制約個室付きのレストランにおいても既に使われているテーブルと新しいテーブルとで場合分けが必要だが，通常の CRP とは異なりさらに大広間か個室かという場合分けも必要となる．そこで以下の五つの場合分けを行った．

(i) 大広間の既に使われているテーブルに座る

(ii) 大広間の新しいテーブルに座る

(iii) 制約個室の既に使われているテーブルに座る

(iv) 制約個室の新しいテーブルに座る

(v) 新しい制約個室の新しいテーブルに座る

この五つの場合分けについて確率割り当ての式を考える．制約 C_k は K に比例して大きくなると仮定し，$C_k = K \times c_k$ となる c_k を導入する．これに応じて二層目のディリクレ分布のパラメータを η から $\frac{\eta}{K}$ に変更する．

(i) 大広間の既に使われているテーブルに座る

はじめに，式 (3) の被積分函数の因子について見てみる．第 1 因子 $P(z_i = (k,j) | \pi, \phi)$ は, 新たに選ばれた単語が第 k 制約の第 j テーブルに座る確率を表し，従って $\pi_k \phi_{kj}$ に等しい．次に第 2 因子は，ベイズの定理と既出単語が多項分布に従うということと上で定めた事前確率の式により，式 (4) が導かれる．

$$\begin{aligned} P(\pi, \phi | z_{1:i-1}, \gamma, \eta, K, C) &\propto P(z_{1:i-1} | \pi, \phi) p(\pi, \phi | \gamma, \eta, K, C) \\ &\propto \left(\prod_{k \notin C} \pi_k^{m_k} \right) \left(\prod_{k \notin C} \pi_k^{\frac{\gamma}{K} - 1} \right) \\ \times \left(\prod_{k \in C} \pi_k^{m_k} \right) \prod_{k \in C} \left\{ \pi_k^{\frac{C_k \gamma}{K} - 1} \left(\prod_{j=1}^{C_k} \phi_{kj}^{m_{kj}} \right) \left(\prod_{j=1}^{C_k} \phi_{kj}^{\eta - 1} \right) \right\} \\ = \prod_{k \notin C} \pi_k^{m_k + \frac{\gamma}{K} - 1} \prod_{k \in C} \pi_k^{m_k + \frac{C_k \gamma}{K} - 1} \prod_{k \in C} \prod_{j=1}^{C_k} \phi_{kj}^{m_{kj} + \eta - 1} \\ \propto \mathrm{Dir}\left(\pi; \left(m_k + \tfrac{\gamma}{K} \right)_{k \notin C}, \left(m_k + \tfrac{|C_k| \gamma}{K} \right)_{k \in C} \right) \\ \times \prod_{k \in C} \mathrm{Dir}(\phi; m_{k1} + \eta, \cdots, m_{kC_k} + \eta) \end{aligned} \quad (4)$$

ここで，既に客が座っているテーブルを選ぶ確率は式 (3) において被積分函数の第 1 因子は $P(z_k = k \mid \pi) = \pi_k$ に帰着し，ϕ に関する積分は 1 となる．このことを踏まえて，式 (3) の右辺を書き直し，$K \to \infty$ とした式を式 (5) に示す.

$$\begin{aligned} &\lim_{K \to \infty} P(z_i = (k,j) \mid z_i^{i-1}, \gamma, \eta, K, C) \\ &= \lim_{K \to \infty} \int \pi_k Dir(\pi; (m_k + \frac{\gamma}{K})_{k \notin C}, (m_k + \frac{C_k \gamma}{K})_{k \in C}) \pi \\ &= \lim_{K \to \infty} \frac{m_k + \frac{\gamma}{K}}{i - i + \gamma} \\ &= \frac{m_k}{i - 1 + \gamma} \end{aligned} \quad (5)$$

これは通常の CRP で既に使われているテーブルに座る確率と一致している.

(ii) 大広間の新しいテーブルに座る

$K_c = \sum_{k \in C} K \cdot c_k$ を用い，K_{i-1} は大広間のテーブルで既に使われているものの個数を表している．そのため，式 (6) 中の $K - K_c - K_{i-1}$ は大広間にあるテーブルの中で選ばれうるテーブルの個数である．中華レストラン過程で新しいテーブルを選択する確率は $\frac{\frac{\gamma}{K}}{i - 1 + \gamma}$ であることを用いると式 (6) となる．

$$\lim_{K\to\infty} \frac{\frac{\gamma}{K}}{i-1+\gamma} \times (K - K_c - K_{i-1})$$
$$= \lim_{K\to\infty} (\frac{\gamma}{i-1+\gamma} \times \frac{K - K_c - K_{i-1}}{K})$$
$$= \lim_{K\to\infty} \frac{\gamma}{i-1+\gamma} \times (1 - \sum_{k\in C} c_k)$$
$$= \frac{\gamma}{i-1+\gamma} \times (1 - \sum_{k\in C} c_k) \quad (6)$$

(iii) 制約個室の既に使われているテーブルに座る

制約個室 k の j 番目のテーブルに座っている人の人数を m_{kj} と表現する．制約個室内のテーブルを選択する場合，ディリクレ森分布に従いディリクレ分布を二回使うため，式 (7) の第一式のようになる．

$$\lim_{K\to\infty} (\int \pi_k Dir(\pi; (m_k + \frac{\gamma}{K})_{k\notin C}, (m_k + \frac{C_k\gamma}{K})_{k\in C}) \times \prod_{l\in C} \phi_{kj} Dir(\phi; m_{l1} + \eta, ..., m_{lC_l} + \eta) d\pi d\phi)$$
$$= \lim_{K\to\infty} (\frac{m_k + \frac{C_k\gamma}{K}}{i-1+\gamma} \times \frac{m_{kj} + \frac{\eta}{K}}{m_k + C_k\frac{\eta}{K}})$$
$$= \lim_{K\to\infty} (\frac{m_k + c_k\gamma}{i-1+\gamma} \times \frac{m_{kj} + \frac{\eta}{K}}{m_k + c_k\eta})$$
$$= \frac{m_k + c_k\gamma}{i-1+\gamma} \times \frac{m_{kj}}{m_k + c_k\eta} \quad (7)$$

(iv) 制約個室の新しいテーブルに座る

制約個室内で新しいテーブルに座るということは既に座っている人数が 0 人のテーブルに座るということになるので，式 (7) の計算において $m_{kj} = 0$ とする．ここで，$K_{k,i-1}$ は制約 k の部屋で既に使われているテーブルの数を表しているため，$C_k - K_{k,i-1}$ は k 番目の個室でまだ使われていないテーブルの数のことである．

$$\lim_{K\to\infty} (\frac{m_k + \frac{C_k\gamma}{K}}{i-1+\gamma} \times \frac{\frac{\eta}{K}}{m_k + C_k\frac{\eta}{K}} \times (C_k - K_{k,i-1}))$$
$$= \lim_{K\to\infty} (\frac{m_k + c_k\gamma}{i-1+\gamma} \times \frac{\frac{\eta}{K}}{m_k + c_k\eta} \times (C_k - K_{k,i-1}))$$
$$= \lim_{K\to\infty} (\frac{m_k + c_k\gamma}{i-1+\gamma} \times \frac{\eta}{m_k + c_k\eta} \times (c_k - \frac{K_{k,i-1}}{K})$$
$$= \frac{m_k + c_k\gamma}{i-1+\gamma} \times \frac{c_k\eta}{m_k + c_k\eta} \quad (8)$$

(v) 新しい制約個室の新しいテーブルに座る

まだ誰も客がいない新しい個室なので $K_{k,i-1} = 0$，また新しいテーブルなので $m_k = 0$ である．これらを式 (8) に代入する．

$$\frac{c_k\gamma}{i-1+\gamma} \times \frac{\eta}{c_k\eta} \times c_k = \frac{c_k\gamma}{i-1+\gamma} \quad (9)$$

これにより以下のようにまとめることができる．

$$P(z_i = (k, j) \mid z_1, ..., z_{i-1})$$
$$= \begin{cases} \frac{m_k}{\gamma + i - 1} & \cdots \text{(i)} \\ \frac{m_k}{\gamma + i - 1} \times \frac{m_{kj} + \eta}{m_k + C_k\eta} & \cdots \text{(ii)} \\ \frac{m_k + c_k\gamma}{i-1+\gamma} \times \frac{m_{kj}}{m_k + c_k\eta} & \cdots \text{(iii)} \\ \frac{m_k + c_k\gamma}{i-1+\gamma} \times \frac{c_k\eta}{m_k + c_k\eta} & \cdots \text{(iv)} \\ \frac{c_k\gamma}{i-1+\gamma} & \cdots \text{(v)} \end{cases}$$

4. 検証

この五つのパターンについての総和をとると，以下の様に確率の和が 1 になっていることがわかる．これにより本研究で求めた計算式の正当性が示された．

$$\sum_{k\notin C} \frac{m_k}{i-1+\gamma} + \frac{\gamma}{i-1+\gamma} \times (1 - \sum_{k\in C} c_k) + \sum_{k\in C} \frac{m_k + C_k\gamma}{i-1+\gamma} \sum_{j=1}^{C_k} \frac{m_k j}{m_k + c_k\eta} + \sum_{k\in C, m_k>0} \frac{m_k + C_k\gamma}{i-1+\gamma} \times \frac{c_k\eta}{m_k + c_k\eta} + \sum_{k\in C, m_k} \frac{c_k\eta}{i-1+\gamma}$$
$$= \sum_{k\notin C} \frac{m_k}{i-1+\gamma} + \frac{\gamma}{i-1+\gamma} \times (1 - \sum_{k\in C} c_k) + \sum_{k\in C} \frac{m_k}{i-1+\gamma} + \sum_{k\in C} \frac{c_k\eta}{i-1+\gamma} = 1 \quad (10)$$

5. おわりに

本研究では潜在的ディリクレ配分法においてトピック数推定の過程である中華レストラン過程に制約知識を組み込んだ．制約知識を組み込むためにディリクレ分布を階層化し，トピック割り当ての確率の計算式を求めた．その際，中華レストラン過程において客が座るテーブルに関して制約知識用の個室とそれ以外の大広間という二つ空間を想定し，五つのパターンに場合分けを行った．五つのパターンについて和を求め 1 となることを確認し，提案した場合分けによる事前確率の正当性を検証した．今後は，ギブスサンプリングの推定で使用できるように計算をし，実際に推定をし，トピック数を推定しながら制約知識を反映出来ているかどうかを確認する．

参考文献

[1] David Andrzejewski, Xiaojin Zhu, and Mark Craven. Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *Proc. of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 25–32, New York, NY, USA, 2009. ACM.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, Vol. 3, pp. 993–1022, mar 2003.

[3] Yuening Hu, Jordan Boyd-Graber, and Brianna Satinoff. Interactive topic modeling. In *Proc. of the 49th ACL-HLT - Volume 1*, pp. 248–257, 2011.

[4] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, Vol. 101, No. 476, pp. 1566–1581, 2006.