3L3-1

居酒屋モデルによるトピックの自発的クラスタリングの実装と実験

Spontaneous Clustering of Topics by Izakaya Model — Implementation and Experiments

立川華代 *1*2　小林一郎 *1　金子晃 *1
Kayo Tatsukawa　Ichiro Kobayashi　Akira Kaneko

*1お茶の水女子大学大学院人間文化創成科学研究科
Ochanomizu University, Graduate School of Humanities and Sciences

Trying to construct the non-parametric version of the Dirichlet forest for the topic model, we actually got a variant of non-parametric LDA model representing the spontaneous clustering of topics. We called this an "Izakaya model" and gave the fundamental formulas for that. Now we present a theoretical formula for its implementation by Gibbs sampling and its realization by concrete probability distribution. We then present experiments using various corpora.

1. はじめに

1.1 問題の動機

統計的言語処理の世界で文書の潜在的意味をトピックと呼ばれるクラスタ推定により解析する LDA ([2]) を用いた研究が活発になり，更にトピック数をも推定するノンパラメトリックなディリクレ過程混合モデルの応用も使われ始めている．我々は Andrejewski ら [1] や Hu ら [4] が扱ったディリクレ森分布を用いた事前制約の取扱いをノンパラメトリック化しようとして，実際にはトピック同士の自発的なクラスタリングを表現する 2 層のディリクレ過程混合モデルに導かれた．これは 2 年前の学会で『居酒屋モデル』の名でその基礎的な公式とともに紹介された ([8])．今回はその続きとしてギッブスサンプリングのための理論式を与え，またを用いた具体的な実験用公式を提案し，それによる実験結果を報告する．紙数の都合で計算の詳細は書けないが，興味を持たれた方はテクニカルレポート [9] をご請求頂きたい．

2. 中華レストラン過程を用いた事前分布の計算

[5] あるいは [7] に書かれた通常の中華レストラン過程を 2 層化し，最初の層では広間に置かれた通常のテーブル，あるいは制約を表す個室が選ばれる．後者の場合は，更にその中のテーブルが別のディリクレ分布で選択されるようにする．

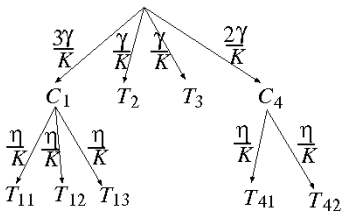

図1．ディリクレ森分布

1 層目の個室への確率割り当てには中のテーブル数だけの重みをつけることで，テーブル毎の重みを二つの層を通して均等にする．こうして K 個のテーブル (クラスタ) 中に C で表される有限な個室 (制約) の集合が含まれるときのディリクレ過程混合モデルの確率を計算する．統計量の記号を m_k は $k \notin C$ のとき広間のテーブル k に着席した客の総数，$k \in C$ のときは個室 k 内の客の総数とし更に後者の場合は m_{kj} でこの個室の第 j テーブルに着いた客の総数とする．以下場合を分けるのを略し広間の場合もテーブル (トピック) を (k, j) で表すことがある．個室 k 内のテーブルの総数を $C_k = Kc_k + o(K)$ として $K \to \infty$ とすると，事前確率として先に [8] 報告した以下の式が得られる：

$$p(z_i = (k,j) | z_1, ..., z_{i-1}) = \begin{cases} \frac{m_k}{i-1+\gamma} \\ \quad \text{(広間のテーブルに客が追加着席)}, \\ \frac{m_k + c_k\gamma}{i-1+\gamma} \frac{m_{kj}}{m_k + c_k\eta} \\ \quad \text{(個室のテーブルに客が追加着席)}, \\ \frac{\gamma}{i-1+\gamma} \times \left(1 - \sum_{k \in C} c_k\right) \\ \quad \text{(広間の新しいテーブルが選ばれた)}, \\ \frac{m_k + c_k\gamma}{i-1+\gamma} \times \frac{c_k\eta}{m_k + c_k\eta} \\ \quad \text{(既存個室の新しいテーブルが選ばれた)}, \\ \frac{c_k\gamma}{i-1+\gamma} \\ \quad \text{(新しい個室のテーブルが選ばれた)}. \end{cases} \quad (1)$$

この結果 n 単語観測後の統計的確率は, テーブルの総数を K_n, そのうち広間のテーブルが $K_{0,n}$ 個, 個室 $k \in C$ のテーブル数を $K_{k,n}$ とすると

$$P(z_{1:n} | \gamma, \eta, C) = \frac{\{\gamma(1 - \sum_{k \in C} c_k)\}^{K_{0,n}} \prod_{k \notin C}^{K_{0,n}} (m_k - 1)!}{(\gamma)_n} \times \prod_{k \in C} \frac{(c_k\gamma)_{m_k} (c_k\eta)^{K_{k,n}} \prod_{l=1}^{C_k} (m_{kl} - 1)!}{(c_k\eta)_{m_k}} \quad (2)$$

となる．ここに $(\alpha)_n := \alpha(\alpha+1)\cdots(\alpha+n-1)$ は上昇階乗である．

連絡先: 金子晃, お茶の水女子大学大学院人間文化創成科学研究科小林研究室, 〒 112-8610 東京都文京区大塚 2-1-1, kaneko.akira@ocha.ac.jp
*2　現在の所属は日本ユニシス株式会社

3. ギッブスサンプリングの公式

上の値は出現順序に依存しないことが証明できるので，ギッブスサンプリングが使える．一般に z_{-i} は $z_{1:n}$ から第 i 要素を除去したものを表すとする．広間のテーブル (トピック)k が語彙を生成する確率分布を $\theta_{(k)}$, 個室のテーブル (k,j) が語彙を生成する確率分布を $\psi_{(kj)}$ とし，Θ は z_{-i} に対応する $\theta_{(k)}$ あるいは $\psi_{(kj)}$ の集合を表すものとすれば，ベイズの定理を用いて

$$\begin{aligned} P(z_i=(k,j)|z_{-i},x_i,\Theta) &= \frac{P(z_i=(k,j),x_i|z_{-i},\Theta)}{P(x_i|z_{-i},\Theta)} \\ &\propto P(z_i=(k,j)|z_{-i})P(x_i|z_i=(k,j),\Theta) \end{aligned} \tag{3}$$

となる．この第 1 因子は (1) において i を n に，m_k, m_{kj} を $m_{-i,k}$, $m_{-i,kj}$ に置き換えたものとなる．また第 2 因子は，分類の順序は上と同じとして

$$\begin{aligned} &p(x_i \mid z_i=(k,j),\Theta) \\ &= \begin{cases} p(x_i \mid \theta_{(k)}), \\ p(x_i \mid \psi_{(kj)}), \\ \displaystyle\int p(x_i \mid \theta) G_0(\theta) d\theta, \\ \displaystyle\int p(x_i \mid \psi, \{\psi_{kj'} \in \Theta\}) G_1(\psi) d\psi, \\ \displaystyle\int p(x_i \mid \psi) G_1(\psi) d\psi \end{cases} \end{aligned} \tag{4}$$

となる．ここで，G_0, G_1 はそれぞれ $\theta_{(k)}$, $\psi_{(kj)}$ の生成規則である．実際に使うのは崩壊 (周辺化) ギッブスサンプリングなので，これらを (3) に代入し両辺に $p(\theta_{(k)|x_{-i}})$ あるいは $p(\psi_{(kj)} \mid x_{-i})$ を掛けて左辺の Θ を積分消去すれば，結局次が得られる：

$$\begin{aligned} &p(z_i=(k,j) \mid z_{-i},x_i,x_{-i}) \\ &\propto \begin{cases} \dfrac{m_{-i,k}}{n-1+\gamma} \dfrac{\int p(x_i|\theta_{(k)}) \prod_{s:s\neq i, z_s=k} p(x_s|\theta_{(k)}) G_0(\theta_{(k)}) d\theta_{(k)}}{\int \prod_{s:s\neq i, z_s=k} p(x_s|\theta_{(k)}) G_0(\theta_{(k)}) d\theta_{(k)}}, \\ \dfrac{m_{-i,k}+c_k\gamma}{n-1+\gamma} \dfrac{m_{-i,kj}}{m_{-i,k}+c_k\eta} \\ \quad \times \dfrac{\int p(x_i|\psi_{(kj)}) \prod_{s:s\neq i, z_s=(k,j)} p(x_s|\psi_{(kj)}) G_0(\psi_{(kj)}) d\psi_{(kj)}}{\int \prod_{s:s\neq i, z_s=(k,j)} p(x_s|\psi_{(kj)}) G_0(\psi_{(kj)}) d\psi_{(kj)}}, \\ \dfrac{\gamma}{n-1+\gamma}\Big(1-\sum_{k\in C} c_k\Big) \displaystyle\int p(x_i|\theta) G_0(\theta) d\theta, \\ \dfrac{m_{-i,k}+c_k\gamma}{n-1+\gamma} \dfrac{c_k\eta}{m_{-i,k}+c_k\eta} \displaystyle\int p(x_i \mid \psi) G_1(\psi) d\psi, \\ \dfrac{c_k\gamma}{n-1+\gamma} \displaystyle\int p(x_i \mid \psi) G_1(\psi) d\psi. \end{cases} \end{aligned} \tag{5}$$

最後に $p(x_i \mid \theta_{(k)})$ 等の分布の具体形を仮定して，(5) を実際に反復実験できる形に書き直すため，$k \notin C$ のときの $p(x_i \mid \theta_{(k)})$, $k \in C$ のときの $p(x_i \mid \psi_{(kj)})$, 及び $G_0(\theta)$, $G_1(\psi)$ を与える必要がある．ここでは $\theta_{(k)}$ が語彙の選択を規定する長さ V の確率ベクトルとして，$p(x_i \mid \theta_{(k)})$ は多項分布 $\theta_{(k)}^{v(x_i)}$, ここに，$\theta_{(k)} = (\theta_{(k)}^1, \ldots, \theta_{(k)}^V)$, $v(x_i)$ は単語 x_i の語彙とした．また $p(x_i \mid \psi_{(kj)})$ も同じ次元の多項分布 $\psi_{(kj)}^{v(x_i)}$ とした．$G_0(\theta \mid \beta)$ は V 次元のディリクレ分布，$G_1(\psi \mid \zeta)$ も V 次元のディリクレ分布とした．結果として，次の具体的なサンプリング公式が得られる．

$$\begin{aligned} &p(z_i=(k,j) \mid z_{-i},x_i,x_{-i}) \\ &\propto \begin{cases} \dfrac{m_{-i,k}}{n-1+\gamma} \dfrac{m_{-i,k}^{v(x_i)}+\beta}{m_{-i,k}+V\beta}, \\ \dfrac{m_{-i,k}+c_k\gamma}{n-1+\gamma} \dfrac{m_{-i,kj}}{m_{-i,k}+c_k\eta} \dfrac{m_{-i,kj}^{v(x_i)}+\zeta}{m_{-i,kj}+V\zeta}, \\ \dfrac{\gamma}{n-1+\gamma}\Big(1-\sum_{k\in C} c_k\Big)\dfrac{1}{V}, \\ \dfrac{m_{-i,k}+c_k\gamma}{n-1+\gamma} \dfrac{c_k\eta}{m_{-i,k}+c_k\eta}\dfrac{1}{V}, \\ \dfrac{c_k\gamma}{n-1+\gamma}\dfrac{1}{V}. \end{cases} \end{aligned} \tag{6}$$

最後の行は，実は使われていない個室のすべてに渡る無限個の行から成るが，実際の実験では，これを 1 行とし，分子の c_k を使われていない個室の全体に渡るこの値の和としてサンプリングを実行し，この行が選ばれたときは，空の個室のうち c_k が最も大きな値の部屋を提供するのが自然である．なお，$c_1 = \cdots = c_{k_0} = c$, $c_{k_0+1} = \cdots = 0$，ただし $k_0 c < 1$, と置いた場合は，個室の数が有限値 k_0 で抑えられており，かつどの部屋も同じ確率で選択される (が，各部屋には制約なしの大広間と同様，無限にテーブルが用意できる) という設定を特別な場合として含んでいる．また $c_k = 0$ と置けば，制約無しの通常のノンパラメトリック LDA となる．

3.1 テストセットパープレクシティ

通常使われている式を我々の場合に合わせて修正した次の式 (以下単に perplexity と言う) により，サンプリングの状況を判定した．

$$\begin{aligned} &\exp\Big(-\frac{1}{n}\sum_{i=1}^{n}\log P(x_i)\Big) \\ &= \exp\Big\{-\frac{1}{n}\Big(\sum_{k=z_i\notin C}\log(P(x_i \mid z_i=k)P(z_i=k)) \\ &\quad + \sum_{(k,j)=z_i\in C}\log(P(x_i \mid z_i=(k,j))P(z_i=(k,j)))\Big)\Big\} \\ &= \exp\Big[-\frac{1}{n}\Big\{\sum_{k=z_i\notin C}\log\Big(\frac{m_{k(x_i)}^v(x_i)+\beta}{m_{k(x_i)}+V\beta}\frac{m_{k(x_i)}}{n+\gamma}\Big) \\ &\quad + \sum_{(k,j)=z_i\in C}\log\Big(\frac{m_{kj(x_i)}^v(x_i)+\zeta}{m_{kj(x_i)}+V\zeta}\frac{m_{kj(x_i)}}{n+\eta}\Big)\Big\}\Big] \end{aligned} \tag{7}$$

4. 実験結果と考察

標準コーパスに対して我々のモデルを適用してみた．言語モデルに適用する場合は，文書毎にこのような統計を行い，トピックは共有するために階層化ディリクレ過程混合モデル HDP [6] を用いなければならない．[6] ではトピックを料理に喩え，文書毎に異なる店のテーブルで集計しているが，我々は上に説明した構造を適用するため，トピックの解釈はテーブルのままとし，文書の違いは客の国籍の違いと解釈して，テーブルの生成消滅は共通で行い，テーブル毎の客の生成確率と客数の統計は国籍(文書) 毎に行うこととした(多国籍居酒屋モデル)．これに伴い，式 (6) の後半の因子には適当に文書番号の添え字 d が付く．またパープレクシティは文書毎に式 (7) で計算したものの相乗平均とした．

perplexity が下がることは事後確率が上がることとほぼ等価である．これはサンプリングにより単調には下がらないが，下がるときだけサンプリングを採用すると偏ったローカルミニマムに落ちてしまうので，サンプリングを最初の何回かそのまま走らせ (いわゆる burn-in) その後は perplexity の極小値とそのときの状態を記録しながらサンプリングを遂行した．

以下は 20_newsgroups/talk.politics.misc/の各文書セットからそれぞれ最初の 100 文書ずつを bag of words としたものから成る 20 文書 (約 75 万語) に stemming と stop words の除去，および頻度 10 以下の語のカットを行ったものに対する提案手法の結果であり，単語数 425328, 語彙数 5529, $\gamma = 0.5$, $\eta = 0.7$, $\beta = 0.5$, $\zeta = 0.5$, $c_k = \frac{1}{4}(\frac{2}{3})^{k-1}$ とした．室構造は η の値に敏感である．最後の結果は各抽出トピックについて頻度順に 10 語ずつ示した．抽出数が少なすぎてクラスタリングの傾向を見るには不十分だが，下の方はサンプル数がかなり少なくなるので，有意な結果を得るにはコーパスをもっと大きくしないといけないようである．

(1) を用いた前処理の一部

```
d,i= 0,0: rooms= [0, 1] tables=[[], [0]]
d,i= 0,1: rooms= [0, 1, 2] tables= [[], [0], [1]]
...
d,i= 0 23 : rooms= [0, 1, 2] tables= [[], [0, 2], [1]]
...
d,i= 0 312 : rooms= [0, 1, 2, 3] tables= [[], [0, 2], [1], [3]]
...
d,i= 19 15285 : rooms= [0, 1, 2, 3, 4, 5] tables= [[11], [0, 2, 4, 5, 8],
[1, 7, 10], [3], [6, 9], [12]]
```

Gibbs サンプリング burn-in

```
Starting test_set_perplexity=3879.029378
1-th preliminary iteration: perplexity=3789.258809
rooms= [0, 1, 2, 3, 4] tables= [[11], [0, 2, 4, 5, 8], [1], [3], [6]]
...
20-th preliminary iteration: perplexity=3357.373102
rooms= [0, 1, 2, 3, 4] tables= [[11, 7, 13], [0, 2, 4, 5], [1], [3], [6]]
```

Gibbs サンプリング 本番

```
1-th iteration: perplexity=3355.738433
rooms= [0, 1, 2, 3, 4] tables= [[11, 13, 7], [0, 2, 4, 5], [1], [3], [6]]
...
500-th iteration: perplexity=3905.505643
rooms= [0, 1, 2, 3, 4, 5, 6, 7, 8] tables= [[11, 13], [0, 2, 4, 5],
[1, 8, 7, 10, 21, 17, 23], [3, 25, 26], [6, 12, 15], [9], [14], [18], [20, 16]]
```

終了時のトピック内容の一部 ($K = 24$)

--- room 0-0 ---	--- room 0-1 ---	--- room 1-0 ---
area: 0.001172	thread: 0.001227	hold: 0.001383
turn: 0.001172	content: 0.001227	12: 0.001237
term: 0.001172	consid: 0.001082	put: 0.001237
qualiti: 0.000959	comment: 0.001082	john: 0.001092
commun: 0.000959	opinion: 0.001082	chanc: 0.001092
differ: 0.000959	origin: 0.001082	move: 0.001092
caus: 0.000959	compat: 0.001082	se: 0.001092
happen: 0.000959	ken: 0.000938	utexa: 0.000946
ti: 0.000959	1991: 0.000938	03: 0.000946
greenbelt: 0.000959	newsserv: 0.000938	free: 0.000946

--- room 1-1 ---	--- room 1-2 ---	--- room 1-3 ---
wayn: 0.001050	line: 0.001527	man: 0.001255
01: 0.001050	provid: 0.001527	peter: 0.001255
engin: 0.001050	toronto: 0.001527	act: 0.001255
poster: 0.001050	dog: 0.001168	number: 0.001255
exactli: 0.001050	usernam: 0.000988	ps: 0.001062
origin: 0.001050	view: 0.000988	dept: 0.001062
uiuc: 0.001050	differ: 0.000988	laboratori: 0.001062
nntp: 0.001050	uniform: 0.000988	sender: 0.001062
plu: 0.001050	drive: 0.000988	child: 0.000869
descart: 0.000889	digit: 0.000988	april: 0.000869

--- room 2-0 ---	--- room 2-1 ---	--- room 2-2 ---
cmu: 0.014609	de: 0.001334	version: 0.001266
cs: 0.014279	easili: 0.001334	direct: 0.001071
srv: 0.009978	definit: 0.001177	interfac: 0.001071
cantaloup: 0.006912	task: 0.001020	49: 0.001071
line: 0.006590	dure: 0.001020	uh: 0.001071
subject: 0.006334	word: 0.001020	degre: 0.001071
messag: 0.006323	02: 0.001020	pass: 0.001071
apr: 0.006261	ingr: 0.001020	bogu: 0.000877
date: 0.005852	deal: 0.001020	attend: 0.000877
newsgroup: 0.005798	rob: 0.001020	light: 0.000877

--- room 2-3 ---	--- room 2-4 ---	--- room 2-5 ---
softwar: 0.001241	experi: 0.001291	organ: 0.001399
book: 0.001241	00: 0.001139	articl: 0.001105
easili: 0.001076	att: 0.001139	step: 0.001105
neglect: 0.001076	sort: 0.001139	start: 0.001105
add: 0.001076	blue: 0.000987	info: 0.001105
line: 0.001076	lead: 0.000987	summari: 0.001105
attent: 0.001076	choic: 0.000987	intend: 0.001105
honor: 0.001076	al: 0.000987	form: 0.000958
hp: 0.000910	mark: 0.000987	creation: 0.000958
brian: 0.000910	parti: 0.000836	decwrl: 0.000958

--- room 2-6 ---	--- room 3-0 ---	--- room 3-1 ---
build: 0.001027	free: 0.001291	effect: 0.001045
august: 0.001027	doug: 0.001093	case: 0.001045
short: 0.001027	import: 0.001093	hard: 0.001045
program: 0.001027	folk: 0.001093	stop: 0.001045
sale: 0.001027	white: 0.001093	product: 0.001045
hear: 0.001027	advanc: 0.001093	stanford: 0.001045
illinoi: 0.000733	rule: 0.000894	offici: 0.000813
stori: 0.000733	make: 0.000894	ucsd: 0.000813
found: 0.000733	note: 0.000894	group: 0.000813
orient: 0.000733	claim: 0.000894	real: 0.000813

--- room 3-2 ---	--- room 4-0 ---	--- room 4-1 ---
common: 0.001049	access: 0.001574	mother: 0.001175
happi: 0.001049	de: 0.001242	43: 0.001175
notic: 0.001049	vax: 0.001242	close: 0.001175
pretti: 0.001049	joe: 0.001077	begin: 0.001018
event: 0.000749	ac: 0.001077	request: 0.001018
howland: 0.000749	eric: 0.001077	zaphod: 0.001018
nonsens: 0.000749	research: 0.000911	alon: 0.001018
florida: 0.000749	laboratori: 0.000911	tue: 0.001018
mp: 0.000749	servic: 0.000911	marbl: 0.001018
testifi: 0.000749	place: 0.000911	assum: 0.001018

--- room 4-2 ---	--- room 5-0 ---	--- room 6-0 ---
austin: 0.001098	gener: 0.001128	newsgroup: 0.001414
occur: 0.001098	elroi: 0.001128	add: 0.001265
mine: 0.001098	softwar: 0.001128	elroi: 0.001265
pc: 0.000898	sourc: 0.001128	coupl: 0.001265
number: 0.000898	titl: 0.001128	200: 0.001116
kind: 0.000898	gui: 0.001128	ufl: 0.001116
model: 0.000898	affect: 0.001128	lab: 0.000967
present: 0.000898	georg: 0.000977	pc: 0.000967
gui: 0.000898	respond: 0.000977	stori: 0.000967
octob: 0.000898	mark: 0.000977	choic: 0.000967

5. まとめと課題

本研究において，中華レストラン過程における事前確率分布にディリクレ森分布を導入し，クラスタに出現するトピック同士の自発的なクラスタリングを表現する枠組みを作った．またギッブスサンプリングの式を導き，実験を通じて，トピックのクラスタ集合が生成消滅する様子を観察できた．トピック同士のクラスタリングは，無限個のパラメータ c_k に強く依存する．今後の課題として，他のハイパーパラメータと並んで，c_k についても推定できるようにするのは興味深いと思われる．

参考文献

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

[2] David M. Blei, Andrew Y. Ng, Michael I. Jordan, and John Lafferty, Latent Dirichlet allocation, *Journal of Machine Learning Research*, **3**(2003), 993–1022.

[3] Griffiths T, Steyvers M., Finding scientific topics, PNAS **101** suppl.1 (2004), 5228–5235.

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

[5] R. M. Neal, Markov chain sampling methods for Dirichlet process mixture models, J. of Computational and Graphical Statistics, **9**-2,(2000), 249–265.

[6] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, Hierarchical Dirichlet processes, *Journal of the American Statistical Association*, **101**(2006),1566–1581.

[7] 上田修功, 山田武士, ノンパラメトリックベイズモデル, 応用数理 **17**-3 (2007), 248–257.

[8] 立川華代，小林一郎，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，人工知能学会 2013 大会予稿 1F3-5.

[9] 立川華代，小林一郎，金子晃，居酒屋モデルによるトピックの自発的クラスタリング— 理論と実験，Technical Report Ocha-IS No.14-1.