D-015

LDA による有意なトピック分析が可能な文書集合の量的な考察
Substantively examination of that is possible for document set to significant topic analysis

古澤 昂典† 冨浦洋一‡
Kosuke Furusawa Yoichi Tomiura

1. はじめに

データベースの発展に伴い，学術論文も電子媒体として管理され，ユーザはタイトルや著者名等で特定されている論文に容易にアクセスすることが可能になった．求める論文がタイトル等で特定されておらず，「～に関する論文」のような内容に基づく情報要求を満たす論文が求める論文である場合はキーワード検索が多く使われる．キーワード検索の問題として「大まかなクエリで検索される膨大な数の論文の内容確認の困難さ」があげられる．近年では対策として，文書集合に対しその文書の話題を表すトピックの分析を行ない，その結果を利用して検索を支援する等の研究が行なわれている[1]．また，キーワードで検索した論文集合から抽出されたトピックにより当該の論文集合を俯瞰的に把握することができ，ユーザは調査対象を要求に合ったトピックを含む論文のみに絞り込むことが出来る．こうしたトピック分析に使われる統計モデルとして LDA が盛んに使用されている[2][3].

本研究では，論文アブストラクトを対象とした Gibbs Sampling によるトピック分析を想定している． Gibbs Sampling による統計モデルのパラメタ推定では，生成した各サンプルから推定されるパラメタ値の標本平均をパラメタの推定値とする．Gibbs Sampling に基づく LDA では，トピックは単なる番号であり，文書毎の各トピックの発生確率を表すパラメタの事前分布はトピックに関して対称である．また，トピック毎の各単語の発生確率を表すパラメタの事前分布もトピックによる違いはない．このため Gibbs Sampling に基づく LDA では，文書集合のサイズ（延べ単語数）が小さい場合，Gibbs Sampling により生成した各サンプルから推定されるパラメタ値の標本平均をパラメタの推定値とすると，有意なトピック分析ができない可能性がある．キーワード検索で得られる論文アブストラクト集合のサイズは，LDA によるトピック分析を用いた研究で報告されている文書集合のサイズに比べ非常に小さい．

そこで，本研究では，LDA が仮定する統計モデル（各パラメタはその事前分布に従ってランダムに設定）から文書集合を生成し，Gibbs Sampling を用いてパラメタ推定を行ない，文書サイズのパラメタ推定の精度への影響を調査し，キーワード検索で得られる論文アブストラクト集合のトピック分析に Gibbs Sampling に基づく LDA が適用できるかどうかを検討する．

†九州大学　大学院システム情報科学府
Kyushu University Graduate School of information Science and Electrical Engineering
‡九州大学　大学院システム情報科学研究院
Kyushu University Faculty of information Science and Electrical Engineering

2. LDA の言語モデルとパラメタ推定

2.1 言語モデル

LDA では文書は単語の系列であるとみなす．各単語は潜在変数であるトピックから生成されるとする． m 番目の文書の i 番目の単語を $w_i^{(m)}$ で表し，文書の単語数を l_m で表す．トピック数を K，単語の異なり数を V とする．文書 m でトピック t を持つ語が生成される確率を $\theta_t^{(m)}$ と表し，$\theta^{(m)} = (\theta_1^{(m)}, \theta_2^{(m)}, \cdots, \theta_K^{(m)})$ と表記する．また，トピック t で語 w が発生する確率を $\phi_w^{(t)}$ と表し，$\phi^{(m)} = (\phi_1^{(t)}, \phi_2^{(t)}, \cdots, \phi_V^{(t)})$ と表記する．m 番目の文書 $\mathbf{w}^{(m)} = (w_1^{(m)}, w_2^{(m)}, \cdots, w_{lm}^{(m)})$ と，各単語に付与されるトピックの列 $\mathbf{z}^{(m)} = (z_1^{(m)}, z_2^{(m)}, \cdots, z_{lm}^{(m)})$ が発生する確率は

$$
\begin{aligned}
&P(\mathbf{w}^{(m)}, \mathbf{z}^{(m)} \mid \theta, \phi) \\
&\quad = \prod_{i=1}^{l_m} P(z_i^{(m)} \mid \theta^{(m)}) P(w_i^{(m)} \mid z_i^{(m)}, \phi) \\
&\quad = \prod_{i=1}^{l_m} \theta_{z_i^{(m)}}^{(m)} \phi_{w_i^{(m)}}^{(z_i^{(m)})}
\end{aligned}
$$

と表される．

M を総文書数とし，$\mathbf{w}$，$\mathbf{z}$ を

$$\mathbf{w} = (\mathbf{w}^{(1)} \cdots \mathbf{w}^{(M)}), \quad \mathbf{z} = (\mathbf{z}^{(1)} \cdots \mathbf{z}^{(M)})$$

とすると，$(\mathbf{w}, \mathbf{z})$ が生成される確率は以下のようになる．

$$
\begin{aligned}
&P(\mathbf{w}, \mathbf{z} \mid \theta, \phi) \\
&\quad = \prod_{m=1}^{M} P(\mathbf{w}^{(m)}, \mathbf{z}^{(m)} \mid \theta, \phi) \\
&\quad = \prod_{m=1}^{M} \prod_{i=1}^{l_m} \theta_{z_i^{(m)}}^{(m)} \phi_{w_i^{(m)}}^{(z_i^{(m)})} \\
&\quad = \prod_{m=1}^{M} \prod_{k=1}^{K} \{\theta_k^{(m)}\}^{n(m,k)} \times \prod_{k=1}^{K} \prod_{w=1}^{V} \{\phi_w^{(k)}\}^{n(k,w)}
\end{aligned}
$$

$\theta^{(m)}$ の事前分布 $\pi(\theta^{(m)}|\alpha)$ をディレクレ分布

$$\frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \prod_{k=1}^{K} \{\theta_k^{(m)}\}^{\alpha-1}$$

とし，$\phi^{(k)}$ の事前分布 $\pi(\phi^{(k)}|\beta)$ をディクレレ分布

$$\frac{\Gamma(V\beta)}{\Gamma(\beta)^V}\prod_{w=1}^{V}\{\phi_w^{(k)}\}^{\beta-1}$$

とする．すると，パラメタ α, β を与えたときの $\mathbf{w},\mathbf{z},\theta,\phi$ の結合確率 (1)が導かれる．

$$\begin{aligned}P(\mathbf{w},\mathbf{z},\theta,\phi|\alpha,\beta)&=P(\mathbf{w},\mathbf{z}|\theta,\phi)\cdot\prod_{m=1}^{M}\pi(\theta^{(m)}|\alpha)\cdot\prod_{k=1}^{K}\pi(\phi^{(k)}|\beta)\\&=\left\{\frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K}\right\}^M\prod_{m=1}^{M}\prod_{k=1}^{K}\left\{\theta_k^{(m)}\right\}^{n(m,k;\mathbf{w},\mathbf{z})+\alpha-1}\\&\quad\times\left\{\frac{\Gamma(V\beta)}{\Gamma(\beta)^V}\right\}^K\prod_{k=1}^{K}\prod_{w=1}^{V}\left\{\phi_w^{(k)}\right\}^{n(k,w;\mathbf{w},\mathbf{z})+\beta-1}\end{aligned}\tag{1}$$

ここで，n(m,k;**w**,**z**)は文書 m 中にトピック k が表れる回数，n(k,w;**w**,**z**)はトピック k で単語 w が表れる回数である．
θ, ϕ を積分消去して，

$$P(\mathbf{w},\mathbf{z}|\alpha,\beta)=\int_{\Theta}\int_{\Phi}P(\mathbf{w},\mathbf{z},\theta,\phi|\alpha,\beta)d\theta d\phi$$

$$P(\mathbf{z}|\mathbf{w},\alpha,\beta)=\frac{P(\mathbf{w},\mathbf{z}|\alpha,\beta)}{\sum_{\mathbf{z}}P(\mathbf{w},\mathbf{z}|\alpha,\beta)}$$

を得る．上記の $P(\mathbf{z}|\mathbf{w},\alpha,\beta)$ に従った **z** を Gibbs Sampling を利用して生成し，次節で述べるようにして θ, ϕ を推定する．

2.2 パラメタ推定

上記の $P(\mathbf{z}|\mathbf{w},\alpha,\beta)$ に従った **z** を Gibbs Sampling を利用して生成する（Gibbs Sampling の方法については省略する）．最初 **z** をランダムに与え，Gibbs Sampling を繰り返す．初期部分で得られるサンプルは初期値の影響があるため捨てて，**z** の分布が定常分布に落ち着いた後のサンプルを利用してパラメタ推定を行なう．得られたサンプルを **z** とすると，(**w**, **z**) を与えたときのθ,ϕの事後分布による$\theta_k^{(m)}$, $\phi_w^{(k)}$ の事後平均値としてパラメタの推定式(2)(3)が得られる．

$$\begin{aligned}\tilde{\theta}_k^{(m)}(\mathbf{z})&=E[\theta_k^{(m)}|\pi(\theta,\phi|\mathbf{w},\mathbf{z},\alpha,\beta)]\\&=\int\int\theta_k^{(m)}\pi(\theta,\phi|\mathbf{w},\mathbf{z},\alpha,\beta)d\theta d\phi\\&=\frac{n(k,m;\mathbf{w},\mathbf{z})+\alpha}{\sum_k n(k,m;\mathbf{w},\mathbf{z})+K\alpha}\end{aligned}\tag{2}$$

$$\begin{aligned}\tilde{\phi}_w^{(k)}(\mathbf{z})&=E[\phi_w^{(k)}|\pi(\theta,\phi|\mathbf{w},\mathbf{z},\alpha,\beta)]\\&=\int\int\phi_w^{(k)}\pi(\theta,\phi|\mathbf{w},\mathbf{z},\alpha,\beta)d\theta d\phi\\&=\frac{n_{zw}(k,w;\mathbf{w},\mathbf{z})+\beta}{\sum_w n_{zw}(k,w;\mathbf{w},\mathbf{z})+V\beta}\end{aligned}\tag{3}$$

しかし，**z** は観測データとして与えられたものではない．文献[3]では，上記の推定式が示されているのみで，どのような **z** に従って推定すればよいのかについては言及されていない．Gibbs Sampling は $P(\mathbf{z}|\mathbf{w},\alpha,\beta)$ に従った **z** を生成するだけであるから，偶然 $P(\mathbf{z}|\mathbf{w},\alpha,\beta)$ の値が小さな **z** を生成する可能性もある．そのような **z** を用いた場合のパラメタの推定値はデータの性質を表すものではない．一方，通常，Gibbs Sampling を利用してパラメタ推定を行なう場合，生成した各サンプルから推定されるパラメタ値の標本平均をパラメタの推定値とする．この方法に従えば，LDA の場合も，以下に導出するように，生成された各サンプル毎のパラメタ推定値の標本平均を取ることで文書データ **w** が与えられたときのパラメタの事後平均として，パラメタ値が推定できる．

$$\begin{aligned}\hat{\theta}_k^{(m)}&=\int\int\theta_k^{(m)}\pi(\theta,\phi|\mathbf{w},\alpha,\beta)d\theta d\phi\\&=\int\int\theta_k^{(m)}\frac{P(\mathbf{w},\theta,\phi|\alpha,\beta)}{P(\mathbf{w}|\alpha,\beta)}d\theta d\phi\\&=\int\int\theta_k^{(m)}\frac{\sum_{\mathbf{z}}P(\mathbf{w},\mathbf{z},\theta,\phi|\alpha,\beta)}{P(\mathbf{w}|\alpha,\beta)}d\theta d\phi\\&=\sum_{\mathbf{z}}\frac{P(\mathbf{w},\mathbf{z}|\alpha,\beta)}{P(\mathbf{w}|\alpha,\beta)}\int\int\theta_k^{(m)}\frac{P(\mathbf{w},\mathbf{z},\theta,\phi|\alpha,\beta)}{P(\mathbf{w},\mathbf{z}|\alpha,\beta)}d\theta d\phi\\&=\sum_{\mathbf{z}}P(\mathbf{z}|\mathbf{w},\alpha,\beta)\int\theta_k^{(m)}\pi(\theta,\phi|\mathbf{w},\mathbf{z},\alpha,\beta)d\theta d\phi\\&=\sum_{\mathbf{z}}P(\mathbf{z}|\mathbf{w},\alpha,\beta)\cdot\tilde{\theta}_k^{(m)}(\mathbf{z})\end{aligned}$$

大数の法則より，

$$\cong\frac{1}{T}\sum_{t=1}^{T}\tilde{\theta}_k^{(m)}(\mathbf{z}_{T_0+t})\tag{4}$$

上記において$\mathbf{z_t}$は t 回目に Gibbs Sampling により得られた **z** のサンプルを表す．$\mathbf{z_1}$から$\mathbf{z_{T_0}}$はランダムに与えた **z** の初期値の影響があるとして破棄する．T はパラメタ推定に十分な大きな数である．
同様にして，

$$\hat{\phi}_w^{(k)}\cong\frac{1}{T}\sum_{t=1}^{T}\tilde{\phi}_w^{(k)}(\mathbf{z}_{T_0+t})\tag{5}$$

１つの文書のサイズ（延べ単語数）が大きく，文書集合のサイズ（延べ単語数）も大きい場合，単一の **z** に基づいたパラメタ推定値でも，上記の標本平均によるパラメタ推定値にごく近い可能性が高い．一方，本研究で対象としているキーワード検索で得られる論文のアブストラクト集合のような文書集合の場合，単一の **z** に基づいたパラメタ推

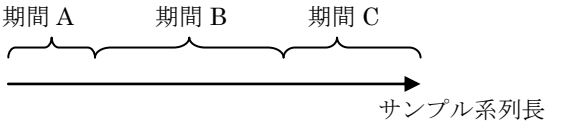

図 1

定では誤差が大きい可能性があり，標本平均に基づくパラメタ推定を行なうべきであると考えられる．ただし，この推定法は，次節で述べるような問題を持っている．

2.3 サンプル系列における互換の頻発とその対処

2.3.1 サンプル系列における互換の頻発

パラメタ推定値の標本平均が文書データ $\mathbf{w}$ が与えられたときのパラメタの事後平均の良い近似であるためには，用いるサンプルの系列の長さ T が十分に大きくなければならない．ここで，LDA で用いている統計モデルは内部変数の実現値であるトピックに関して対称であるので，$\sigma(\mathbf{z})$ を $\mathbf{z}$ の互換（例えば，$\mathbf{z}$ 中のトピック 1 をすべてトピック 2 で置換し，$\mathbf{z}$ 中の 2 をすべて 3 で置換，$\mathbf{z}$ 中の 3 をすべて 1 で置換したもの）とすると

$$P(\mathbf{z}\mid\mathbf{w},\alpha,\beta)=P(\sigma(\mathbf{z})\mid\mathbf{w},\alpha,\beta)$$

となる．これは，式(1)から明らかであろう．このため，推定に用いるサンプルの系列を十分に長く取ると，サンプル $\mathbf{z}$ とそのある互換がサンプルの系列中でほぼ同数出現すると考えられる．本論文では，この現象を「サンプル系列における互換の頻発」と呼ぶ．

そのような長い系列を用いてパラメタ推定値の標本平均を求めると，トピックに関してパラメタが表す確率分布が均一化した以下の様な特徴のないものになってしまうと予想される．

$$\theta_k^{(m)}=\frac{1}{K}\quad(\forall m,\forall k),$$

$$\phi_w^{(k)}=\phi_w^{(k')}\ (\forall k,\forall k')$$

観測データ（文書である単語列の系列）の長さと個々の単語のトピックを指定した内部変数の実現値の系列 $\mathbf{z}$ の長さは等しい．観測データが長い系列であればあるほど，ある時点で内部変数の系列の実現値として $\mathbf{z}$ をサンプリングしたとき，以降のある時点でその互換である$\sigma(\mathbf{z})$ をサンプリングする確率は小さくなり，サンプル系列における互換の頻発は起きにくくなると考えられる．逆に観測データが短い系列の場合，サンプル系列における互換の頻発の問題が深刻になる．

2.3.2 サンプル系列における互換の頻発への対処法の提案

この対処法としてサンプル系列中で，$P(\mathbf{z}\mid\mathbf{w},\alpha,\beta)$ が最も高くなる $\mathbf{z}_{\max}$ を保存しておき，サンプリング終了後に，$\theta_k^{(m)}$，$\phi_w^{(k)}$をそれぞれ

$$\tilde{\theta}_k^{(m)}(\mathbf{z}_{\max}),\quad \tilde{\phi}_w^{(k)}(\mathbf{z}_{\max}) \qquad (6)$$
$$\left(\mathbf{z}_{\max}=\arg\max_{t\in\{1,2,\cdots T\}} P(\mathbf{z}_{T_0+t}\mid\mathbf{w},\alpha,\beta)\right)$$

と推定する方法が考えられる．本論文では，式(6)による推定法を暫定的最大確率の $\mathbf{z}$ に基づくパラメタ推定と呼ぶ．Gibbs Sampling の性質として，$P(\mathbf{z}\mid\mathbf{w},\alpha,\beta)$ が高い $\mathbf{z}$ がサンプリングされる可能性が高いため， Gibbs Sampling を利用して効率的に$P(\mathbf{z}|\mathbf{w},\alpha,\beta)$を最大にする $\mathbf{z}$ の近似値が得られる．

3. 実験

3.1 実験目的

本実験ではまず，サンプル系列における互換の頻発の発生を実験的に確かめる．その後，平均標本に基づくパラメタ推定法と暫定的最大確率の $\mathbf{z}$ に基づくパラメタ推定法のどちらが，文書発生源のパラメタと近いかを確認する．

3.2 パラメタ推定に対する文書サイズ，サンプル系列長の影響と予想

前述のように，どのようなサイズの文書集合に対しても，サンプル系列を十分に長くとるとサンプル系列における互換の頻発が起こり，またサイズの小さな(延べ単語数の少ない)文書集合に関して，暫定的最大確率の $\mathbf{z}$ に基づくパラメタ推定が有効であるということから，サンプル系列と文書サイズに関しての以下の予想を立て検証する(図 1).

(1) 期間 A は，暫定的最大確率の z に基づくパラメタ推定の方が望ましい結果を得られる期間である．期間 A では十分にサンプルを得られないため，標本平均で，観測データが与えられたときのパラメタの事後平均を近似することが出来ないためである．

(2) 期間 B では，サンプル系列を十分取れたことにより，標本平均に基づくパラメタ推定が有効であると考えられる．

(3) 期間 C では，どのような長い文書サイズに対してもサンプル系列を十分に取ると互換の頻発が起こることから，ある程度のサンプル系列をとった後は，暫定的最大確率の $\mathbf{z}$ に基づくパラメタ推定の方が有効であると考えられる．

文書サイズが小さくなるにつれて，サンプル系列における互換の頻発が起こり，期間 B は短くなると予想される．さらに文書サイズが小さくなると，どのようなサンプル系列の長さでも暫定的最大確率の $\mathbf{z}$ によるパラメタ推定の方が有効であると考えられる．

反対に文書サイズが大きくなるにつれ，互換の頻発は発生しにくくなる．文書サイズの増大により互換の頻発が抑制されることから，期間 B は長くなると予想される．

3.3 サンプル系列における互換の頻発の判定

均一化されたパラメタと推定パラメタとの相違度を二乗誤差を用いて計測し，これにより，互換の頻発を判定する．二乗誤差が小さいほど，互換の頻発が起こったこととなる．θとϕについて均一化されたパラメタと推定パラメタの二乗誤差を取った値を U_θ, U_ϕで表す．$\hat{\theta}$と$\hat{\phi}$は標本平均に基づくパラメタ推定値である．

$$U_\theta = \sum_m \sum_k \frac{1}{M * K} (\hat{\theta}_k^{(m)} - \frac{1}{K})^2$$

$$U_\phi = \sum_k \sum_v \frac{1}{K * V} * (\hat{\phi}_w^{(k)} - \bar{\phi}_w)^2$$

ただし，$\bar{\phi}_w = \frac{1}{K} \sum_{k=1}^{K} \hat{\phi}_w^{(k)}$

U_θとU_ϕの和を取り，

$$U = U_\theta + U_\phi \qquad (7)$$

これを，最終的な指標とする．

3.4 実験手順

(1) 事前分布のパラメタ α, β をともに 0.1 として，ディレクレ分布に従ってパラメタθ，ϕを生成し，それらを用いて異なる文書量のトピック付きランダム文書集合を生成する．各文書集合から，発生源のパラメタθ，ϕの最尤推定値を求めておく．また，これらの文書集合のトピックを隠したデータを，LDA の分析対象とする．

(2) それぞれの文書集合に対して，５０万回のサンプリング終了後，各回のパラメタ推定値，その標本平均値を求め，式(7)用いて，サンプル系列における互換の頻発を確認する．これを，第１の実験とする．α，βの値は，文書データを生成したθ，ϕを生成するのに用いた値(α，βともに 0.1)とした．尚，サンプリング開始から 1 万回目までのサンプルは，初期値の影響を強く受けるものとして破棄する．（つまりT_0=10000）

(3) それぞれの文書集合に対し，T_0(=10000)+1 回から，ある回数までサンプリングを行う．その時点での標本平均を用いたパラメタ推定値と暫定的最大確率の $\mathbf{z}$ に基づくパラメタ推定値を求め，どちらがトピック付きの文書集合から最尤推定で得られるパラメタ（発生時のパラメタ）に近いかを比較する．これを第２の実験とする．ただし，発生源のトピック k が推定結果のどのトピックに対応しているかは分からない．このため，σを{1, 2, …, K}上の互換の一つとし，（k が $\sigma(k)$ に対応）$\sigma(\theta^{(m)}) = (\theta_{\sigma(1)}^{(m)}, \theta_{\sigma(2)}^{(m)}, \cdots, \theta_{\sigma(K)}^{(m)})$とすると，発生時のパラメタ$(\theta, \phi)$ と推定されたパラメタ$(\hat{\theta}, \hat{\phi})$の距離を JS-ダイバージェンス D を用いた下式で判定する

$$\min_\sigma \left\{ \frac{1}{K} \sum_z D(\phi^{(z)} \| \hat{\phi}^{(\sigma(z))}) + \frac{1}{M} \sum_m D(\theta^{(m)} \| \sigma(\hat{\theta}^{(m)}) \right\}$$

以下のようなグループで分けられるランダムな文書群を生成し，これを分析する(表 1).

M は文書数，W は１つの文書の単語数，T はトピック数である（延べ単語数は M*W）．

グループ１： 同じ文書数で，１つの文書のサイズが異なる文書集合のグループ．
グループ２： １つの文書のサイズは同じで，文書数が異なる文書集合のグループ．
グループ３： 文書数と１つの文書のサイズの比率が等しいような文書集合のグループ．

表 1 において，異なるグループに関して同じ行で比較すると，延べ単語数が（ほぼ）等しいようにしている．

3.5 実験結果

3.5.1 第 1 実験

実験結果を，表 2,3,4 に示す．全体的に，述べ単語数の増加に従って U の値は大きくなり，互換の頻発が抑えられていくことが分かるが，１つの文書の単語長が短い場合は，同じ延べ単語数に対しても U の値が低い傾向にあることが確認された．これは，1 つの文書サイズが小さければ，θについてのパラメタの均一化が早い段階で進み，θ の影響を受けるϕについても有意なパラメタ推定が出来なくなるためだと考えられる．

表１　実験に使用する文書集合

グループ１	グループ２	グループ３
M30 W10 T5	(M30 W10 T5)	(M30 W10 T5)
M30 W20 T5	M 60 W10 T5	M43 W14 T5
M30 W40 T5	M120 W10 T5	M60 W20 T5
M30 W 80 T5	M240 W10 T5	M80W30T5

表２　グループ１の第１実験結果

	M30W10	M30W20	M30W40	M30W80
U_ϕ	0.0021	0.0062	0.0024	0.0012
U_θ	0.0404	0.0880	0.1141	0.1109
U	0.0425	0.0942	0.1165	0.1121

表 3　グループ 2 の第 1 実験結果

	(M30W10)	M60W10	M120W10	M240W10
U_ϕ	0.0021	0.001	0.001	0.002
U_θ	0.0404	0.040	0.041	0.127
U	0.0425	0.041	0.042	0.129

表 4　グループ 3 の第 1 実験結果

	(M30W10)	M43W14	M60W20	M80W30
U_ϕ	0.0021	0.0010	0.0014	0.0012
U_θ	0.0404	0.0406	0.0630	0.1124
U	0.0425	0.0416	0.0644	0.1136

表 5　M30W10 の文書集合の第 2 実験結果

step	10,000	100,000	250,000	500,000
Average	0.6565	0.7989	0.8030	0.8027
Max_p	0.6565	0.4791	0.6095	0.6094
uniformity	3.7622	3.762	3.7621	3.7621

表 6　M30W80 の文書集合の第 2 実験結果

step	10,000	100,000	250,000	500,000
Average	0.4128	0.3629	0.3632	0.3629
Max_p	0.4128	0.3913	0.3775	0.3893
uniformity	3.2780	3.2780	3.2780	3.2780

表 7　M240W10 の文書集合の第 2 実験結果

step	10,000	100,000	250,000	500,000
Average	0.8074	0.6919	0.3632	0.7454
Max_p	0.8074	0.7011	0.3775	0.7011
uniformity	13.5367	13.5367	13.5367	13.5367

表 8　M80W30 の文書集合の第 2 実験結果

step	10,000	100,000	250,000	500,000
Average	0.4884	0.4249	0.2484	0.4246
Max_p	0.4884	0.4793	0.4738	0.4737
uniformity	4.7781	4.7781	4.7781	4.7781

3.5.2 第 2 実験

標本平均に基づくパラメタ推定に関して，表 5,6,7,8 の実験結果から，サンプル系列が短いうちは推定誤差が大きく，長くなるにつれて，推定に必要な十分なサンプルを得られるため，推定誤差が小さくなり，やがて互換の頻発が起こるため，推定誤差が大きくなるといった当初の予想の傾向が全体的に見られた．

暫定的最大確率となる $\mathbf{z}$ に基づくパラメタ推定についても，全体的な傾向として，サンプル系列の短い間は，標本平均に基づくパラメタ推定より発生源に近い推定になる．サンプリングが長くなるにつれ，十分なサンプル系列を得た標本平均による推定より推定誤差が高くなるが，標本平均によるパラメタが均一化されると，その推定誤差よりは低くなるといった傾向が見られた．本来，暫定的最大確率となる $\mathbf{z}$ に基づくパラメタ推定は安定的に推定誤差が低くなるはずである．この傾向が見られないのは十分に $P(\mathbf{z}\,|\,\mathbf{w}, \alpha, \beta)$ が高い $\mathbf{z}$ が得られていないためと推察できる．

同じ文書量で，二つの推定法の優劣の推移に違いがみられた．暫定的最大確率となる $\mathbf{z}$ に基づくパラメタ推定の方が望ましいのは文書数 30 単語長 10 の文書集合のみである．これは 1 つの文書のサイズが小さい為，サンプル系列における互換の頻発が発生しているからだと考えられる．

以上から，3.2 節での予想の正しさについて，裏付けを得ることが出来たと考えられる．

4. 考察

本研究が対象としている論文アブストラクト集合のサイズに対して，実験の範囲では，少なくともごく少ないサンプル系列しかとらないという場合を除き，標本平均に基づくパラメタ推定の方が望ましいと考えられる．標本平均に基づくパラメタ推定法で，用いるサンプルの系列は長い方が望ましい．一方，サンプル系列が長すぎると，サンプル系列における互換の頻発の問題が出てくるが，実際に分析に用いる文書集合について，どれほどのサンプルを用いて標本平均に基づくパラメタ推定を行うのが望ましいかは不明であるため，互換の頻発を判定し，その前にサンプリングを止めることができるような指標を求めることが今後の課題として考えられる．

5. おわりに

本論文では，まずランダム生成した異なる文書量の文書集合に対して，サンプル系列における互換の頻発が発生するという予測の正しさを確認した．その後，文書集合のサイズ（延べ単語数）が大きいと，サンプル系列がある程度長くとも，標本平均に基づく手法が発生源のパラメタ値と近く，観測データが小さいと，サンプル系列の長さとほぼ関係なく暫定的最大確率となる $\mathbf{z}$ に基づく手法が発生源のパラメタ値と近いことを確認した．結果として，キーワード検索で得た論文アブストラクト集合に対しては，標本平均に基づくパラメタ推定の方が望ましいことを確認できた．

実際のアブストラクトのトピック分析では，サンプル系列と文章サイズの他に，トピック数とアブストラクト中に現れる単語の種類数もパラメタ推定に影響を与えると思われるので，さらに実際のトピックに近い形での実証することが今後の課題としてあげられる．

参考文献

[1] 原島　純，黒橋　禎夫:テキストの表層情報と潜在情報を利用した適合性フィードバック，自然言語処理 19(3):p121-142,2012-09，言語処理学会

[2] D．Blei，A．Ng，and M．Jordan：Latent dirichlet allocation，*the Journal of machine Learning research*，vol．3，pp．993–1022，(2003)

[3] T．L．Griffiths and M．Steyvers：Finding scientific topics．，*Proceedings of the National Academy of Sciences of the United States of America*，vol．101 Suppl 1，pp．5228–35，(2004)