E-019

LDA を用いたトピック分析におけるトピックの理解容易性
Comprehensibility of Topics in Topic Analysis using LDA.

末次 展章† 冨浦 洋一‡
Nobuaki Suetsugu 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．これに対処するために，キーワード検索で得られた結果の論文集合（アブストラクト集合）に対してトピック分析を行い，それを利用して検索の支援を行うことが考えられる．トピック分析に使われる統計モデルとしては，LDA（Latent Dirichlet Allocation）[1][2]が良く使用されている．原島らは，LDA を用いたトピック分析の結果も利用した適合性フィードバックにより文書検索の支援を行う研究を行っており[3]，論文検索の際もこのような手法による支援も考えられる．また，キーワード検索で得られた論文集合のトピック分析の結果から，利用者が要求に合った（興味ある）トピックを指定し，これを含む論文のみに絞り込むことが考えられる．このためには，利用者はトピックが何を表すかが理解できなければならない．しかし，LDA を用いた従来のトピック分析では，トピック k から生成される出現頻度上位の単語からトピック k の内容を解釈しなければならず，利用者にとって分かりづらいものであった．また LDA ではトピックは語の出現分布を規定する潜在的なカテゴリーとして扱われているため，LDA による分析で得られるトピックの中には，我々が通常の意味でトピックと感じるものとは異なるもの（これを本研究では「無意味なトピック」と呼ぶ）も存在し，それが利用者の理解の妨げとなっていた．

我々は，キーワード検索で得られたアブストラクト集合をトピック分析し，その結果を利用者に提示して興味あるトピックを指定させ，そのトピックを含むものだけに絞り込む支援を想定している．本稿では，トピックの理解容易性と相関がある指標を提案する．相関が高ければ，それを利用して解釈が容易な順番にトピックを提示したり，提示するトピックを解釈が容易なものに絞り込んだりすることができる．

2. LDA を用いたトピック分析

†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

LDA[1][2]は，文書生成の確率モデルの一つである．LDA では，各文書がトピックの混合比を持ち，各トピックが単語の出現確率分布を持つとする．そして，文書の生成では，文書の各単語位置に対して，文書のトピックの混合比に従ってトピックが生成され，そのトピックが持つ単語の出現確率分布に従ってその位置の単語が生成されると考える．以下に，本研究で用いる文献[2]の LDA の言語モデルおよび Gibbs Sampling を用いたパラメタ推定法について述べる．

2.1 LDA の言語モデル

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．トピック数を K，単語の異なり数を V とする．また文書 m でトピック k を持つ語が生成される確率を $\theta_k^{(m)}$ とし，トピック k で語 w が発生する確率を $\phi_w^{(k)}$ と表し，

$$\theta^{(m)} = (\theta_1^{(m)}, \theta_2^{(m)}, \cdots, \theta_K^{(m)})$$

$$\phi^{(k)} = (\phi_1^{(k)}, \phi_2^{(k)}, \cdots, \phi_V^{(k)})$$

と表す．m 番目の文書の各単語列を

$$w^{(m)} = (w_1^{(m)}, w_2^{(m)}, \cdots, w_{l_m}^{(m)})$$

とし，各単語に付与されるトピック列

$$z^{(m)} = (z_1^{(m)}, z_2^{(m)}, \cdots, z_{l_m}^{(m)})$$

とすると，$(\boldsymbol{w}^{(m)}, \boldsymbol{z}^{(m)})$ が発生する確率は

$$P(w^{(m)}, z^{(m)} \mid \theta, \phi) = \prod_{i=1}^{l_m} \theta_{z_i^{(m)}}^{(m)} \cdot \phi_{w_i^{(m)}}^{(z_i^{(m)})}$$

と表される．

M を総文書数とし，$\boldsymbol{w}$，$\boldsymbol{z}$ を

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

とすると，$(\boldsymbol{w}, \boldsymbol{z})$が生成される確率は

$$P(\boldsymbol{w}, \boldsymbol{z} \mid \theta, \phi) = \prod_{m=1}^{M} P(w^{(m)}, z^{(m)} \mid \theta, \phi)$$

$$= \prod_{m=1}^{M} \prod_{k=1}^{K} \{\theta_k^{(m)}\}^{n_{DZ}(m,k;\boldsymbol{w},\boldsymbol{z})} \times \prod_{k=1}^{K} \prod_{w=1}^{V} \{\phi_w^{(k)}\}^{n_{ZW}(k,w;\boldsymbol{w},\boldsymbol{z})}$$

と表される．ここで $n_{DZ}(m, k ; \boldsymbol{w}, \boldsymbol{z})$ は $(\boldsymbol{w}, \boldsymbol{z})$ における文書 m 中のトピック k の出現回数であり，$n_{ZW}(k, w ; \boldsymbol{w}, \boldsymbol{z})$ は$(\boldsymbol{w}, \boldsymbol{z})$ におけるトピック k での単語 w の出現回数である．

2.2 Gibbs Sampling とパラメタ推定

文献[2]の LDA では，$\theta^{(m)}, \phi^{(k)}$の事前分布を，次のようにディレクレ分布で与えている．

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

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

パラメタα，βを与えたときの $\boldsymbol{w}$，$\boldsymbol{z}$，θ，ϕの結合確率密度は，

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

となる．θ，ϕを積分消去して，

$$P(\boldsymbol{w}, \boldsymbol{z} \mid \alpha, \beta) = \iint_{\Theta\Phi} P(\boldsymbol{w}, \boldsymbol{z}, \theta, \phi \mid \alpha, \beta) d\theta d\phi$$

を得る．文書集合 $\boldsymbol{w}$ が与えられたときの，各単語に付与された（各単語を生成した）トピックの系列が $\boldsymbol{z}$ である確率は，

$$P(\boldsymbol{z} \mid \boldsymbol{w}, \alpha, \beta) = \frac{P(\boldsymbol{w}, \boldsymbol{z} \mid \alpha, \beta)}{\sum_{\boldsymbol{z}} P(\boldsymbol{w}, \boldsymbol{z} \mid \alpha, \beta)} \quad (1)$$

となる．

トピック分析を行う際に与えられるのは $\boldsymbol{w}$（およびパラメタ α，β，トピック数 K）だけである．文献[2]では，Gibbs Sampling により，式(1)の $P(\boldsymbol{z} \mid \boldsymbol{w}, \alpha, \beta)$ に従ったサンプル $\boldsymbol{z}$ を生成し，生成した $\boldsymbol{z}$ に基づいてパラメタ$\theta^{(m)}, \phi^{(k)}$を$(\boldsymbol{w}, \boldsymbol{z})$ が与えられたときのパラメタの事後分布

$$\pi(\theta, \phi \mid \boldsymbol{w}, \boldsymbol{z}, \alpha, \beta) = \frac{P(\boldsymbol{w}, \boldsymbol{z}, \theta, \phi \mid \alpha, \beta)}{P(\boldsymbol{w}, \boldsymbol{z} \mid \alpha, \beta)}$$

に関する期待値として以下のように推定する．

$$\begin{aligned} \tilde{\theta}_k^{(m)}(\boldsymbol{z}) &= \iint \theta_k^{(m)} \pi(\theta, \phi \mid \boldsymbol{w}, \boldsymbol{z}, \alpha, \beta) d\theta d\phi \\ &= \frac{n_{DZ}(m, k; \boldsymbol{w}, \boldsymbol{z}) + \alpha}{\sum_{k=1}^{K} \{n_{DZ}(m, k; \boldsymbol{w}, \boldsymbol{z}) + \alpha\}} \\ \tilde{\phi}_w^{(k)}(\boldsymbol{z}) &= \iint \phi_w^{(k)} \pi(\theta, \phi \mid \boldsymbol{w}, \boldsymbol{z}, \alpha, \beta) d\theta d\phi \\ &= \frac{n_{ZW}(k, w; \boldsymbol{w}, \boldsymbol{z}) + \beta}{\sum_{k=1}^{K} \{n_{ZW}(k, w; \boldsymbol{w}, \boldsymbol{z}) + \beta\}} \end{aligned}$$

文献[2]で示されているパラメタ推定は上記の通りであるが，論文アブストラクトは 1 文書が比較的小さく，また，想定している文書数は数千である．このため上記の推定式では，最終的にどのサンプル $\boldsymbol{z}$ を用いて推定するかで結果が異なってくる．そこで，本研究では，以下のように上記の推定値の標本平均をパラメタ推定値とする（詳しくは文献[4]を参照）．

$$\hat{\theta}_k^{(m)} \cong \frac{1}{T} \sum_{t=1}^{T} \tilde{\theta}_k^{(m)}(\boldsymbol{z}(T_0 + t))$$

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

ただし，$\boldsymbol{z}(t)$ は t 回目のサンプル（文書集合の各単語に付与されたトピックの系列）を示す．$\boldsymbol{z}(1)$ から $\boldsymbol{z}(T_0)$ までは，初期値の影響があるものとして破棄する．

3. これまでの研究と課題

我々は，キーワード検索で得られるアブストラクト集合をトピック分析し，その結果（トピック）を提示して，検索者が興味あるトピックを指定し，そのトピックを含むアブストラクトに絞り込むという検索支援を想定している．このためには，トピックが何を表しているかを利用者が解釈できなければならない．

ところが，キーワード検索で得られる学術論文のアブストラクト集合を LDA を用いてトピック分析した場合，トピック k 毎に$\phi_w^{(k)}$ の値が高い語 w を提示しただけではトピックの解釈が困難な場合がほとんどであった．このため，これまで，トピックを代表するような語だけを提示する手法の開発や言語モデルを改良して複合的な表現の解析も同時に行なうようなモデルの構築を行なってきた．これらはある程度効果はあったものの，依然として解釈が困難なトピックがあった．LDA ではトピックは語の出現分布を規定する潜在的なカテゴリーとして扱われているため，LDA による分析で得られるトピックの中には，我々が通常の意味でトピックと感じるものとは異なる「無意味なトピック」も存在し，これらのトピックに対してはトピックを代表する語を提示したり提示する語に複合表現も含めたりしても，トピックの解釈を助けることにはならない．

そこで，本研究では，最終的なトピック数よりある程度大きなトピック数で LDA を用いたトピック分析を行ない，できるだけ解釈が容易なトピックだけを提示することを目的とする．ただし，解釈容易性は，解釈する人の背景知識等に依存するため，個人差がある．しかも，解釈容易性は実際に解釈してみなければ測定できないため，LDA による分析結果を利用して定量化できる量で，解釈容易性と相関がある指標を提案する．

4. 提案手法

まず，文書集合から 1 つの文書を選ぶことを考える．各文書を等確率で選ぶ場合は，選ばれる文書に関する情報量は，

$$-\sum_{m=1}^{M}\frac{1}{M}\log\frac{1}{M}=\log M$$

である．一方，トピック k を指定し，トピック k が含まれる割合，つまり，確率

$$p(m\,|\,k)=\frac{\theta_k^{(m)}}{\sum_m \theta_k^{(m)}}$$

に従って文書を選ぶ場合，選ばれる文書に関する情報量は，

$$H(k)=-\sum_m p(m\,|\,k)\log p(m\,|\,k)$$

である．トピック k を指定することで，選ばれる文書に関する情報量（乱雑さ）が，

$$\log M - H(k) \qquad (2)$$

だけ減少するため，本稿では，式(2)を「文書の特定に対するトピック k の相互情報量」と呼ぶ．この相互情報量が低いトピックは，様々な文書に散在することになり，そのため意味的なまとまりが悪く解釈が困難である可能性が高いか，意味的なまとまりがあったとしても，キーワードによる検索結果の文書を絞り込むという今回のトピック分析の利用法からして，役に立たないトピックである．

そこで，LDA によるトピック分析の結果に対して，各トピックに対して，式(2)の値を求め，この値が大きなものだけを対象として，トピックの提示を行なう．

5. 実験

5.1 実験の目的

本実験の目的は文書の特定に対するトピックの相互情報量とトピックの解釈容易性に大きな相関があることを確認し，もしそうであるならば，どこまでのトピックを利用者に提示すれば良いかを検討することである．

5.2 実験手順

トピック分析のプログラムは，GibbsLDA++[5] に手を加えたものを用いた．実験手順を以下に示す．

(1) LDA によるトピック分析

複数のメタパラメタ（図 1）を用いて LDA によるトピック分析を行う．この際使用した実験データはキーワード「電波伝搬」で検索された論文アブストラクト 1078 文書，異なり単語数 2307 語である．GibbsLDA++では α のデフォルト値は，$K/50$ であり，今回の実験ではトピック数 K は 50 であるため，$\alpha = 1$ となる．しかし，今回の実験では論文アブストラクトを対象としており，比較的文書が短いため，$\alpha \leq 1.0$ で実験を行った．またこれまでに行った AIC を基準として文書集合毎の適切なトピック数を求めるという実験において 1,000 文書程度の論文集合ならば T = 25~35 程度であったため，それよりも大きなトピック数 50 で実験した．なお，2.2 節で示したように，今回の実験では Gibbs Sampling により生成されたサンプルによるパラメタ推定値の標本平均でパラメタを推定した．対数尤度の収束具合から，5,000 回目までで十分に定常状態になっていると考えられたため，T_0=5,000 とし，サンプリング 5001 回目から 10,000 回目までの各回におけるパラメタ推定値の標本平均をとって $\hat{\theta}$，$\hat{\phi}$ を計算した．

トピック数	α	β
50	0.1	0.1
50	0.2	0.1
50	0.5	0.1
50	1.0	0.1

図 1:実験に使用したパラメタ

(2) 各トピックの被験者による解釈

LDA の出力結果（トピック k 毎の $\phi_w^{(k)}$ の上位 50 語までの語 w のリスト）を見て，解釈容易性を 4 段階で評価する．

(3) 各トピックの相互情報量と解釈容易性の関係調査

文書の特定に対するトピックの相互情報量（式(2)の値）でトピックをソートし，上記で求めた解釈容易性との関係性について調査する．また相互情報量と解釈容易性の評価値との間の相関係数を求める．

5.3 実験結果と考察

次に示す 5 つのグラフ（図 2-図 5）は各トピックを俯瞰し，解釈容易性を判定した後，各トピックの相互情報量を求め降順にソートしたものである．線グラフは各トピックの相互情報量を表している．棒グラフを各トピックの解釈容易性を表している(0.5，1.0，1.5，2.0)．今回の実験では 0.5，1.0 と評価されたトピックを解釈が困難なトピックとしている．

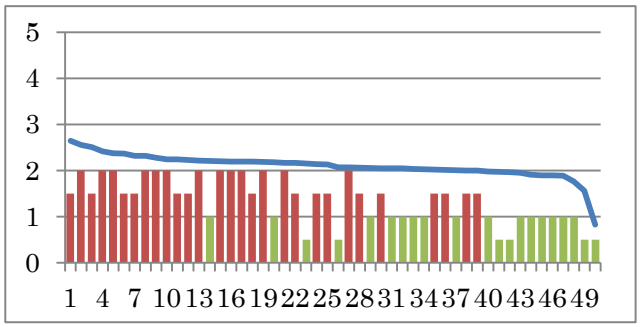

図 2:相互情報量と各トピックの解釈容易性評価値(α=0.1)

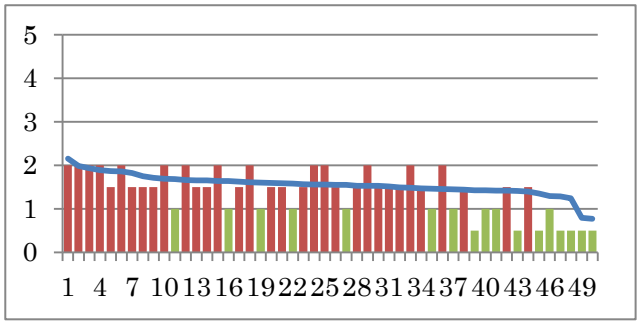

図 3:相互情報量と各トピックの解釈容易性評価値(α=0.2)

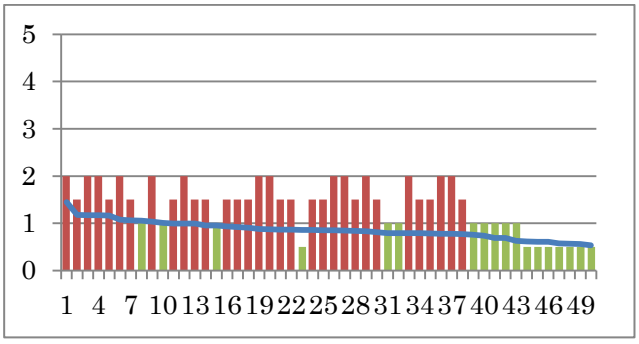

図 4:相互情報量と各トピックの解釈容易性評価値(α=0.5)

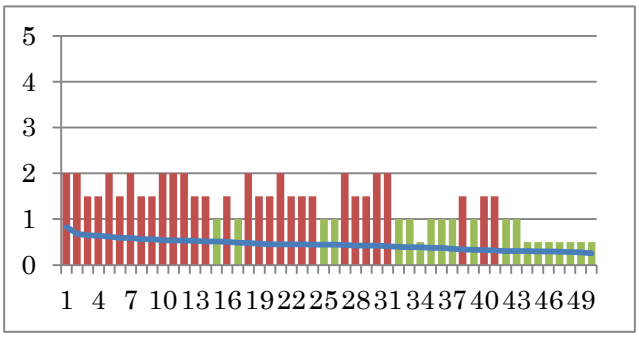

図 5:相互情報量と各トピックの解釈容易性評価値(α=1.0)

次にパラメタ毎の相互情報量と解釈容易性の評価値との間の相関係数を図 7 に示す.

α	0.01	0.1	0.2	0.5	1.0
相関係数	0.324	0.309	0.321	0.265	0.239

図 6:相互情報量と解釈容易性評価値の相関係数

相互情報量とトピックの関係性の調査実験より，相互情報量が大きなトピックは理解容易なものが多いという傾向が得られた．よって相互情報量が大きいトピックから順に提示することで検索者に理解容易なトピックだけを提示することが可能になると考えられる．しかし相互情報量が大きいにも関わらず理解困難なトピックや，相互情報量が小さいにも関わらず理解容易なトピックも存在した． これは今回の評価者には背景知識の不足のため解釈できなかったが背景知識を持った別の評価者であれば解釈できる可能性がある.

また相互情報量と解釈容易性評価値の相関係数を見ると弱い相関がみられた．このことからもトピックの相互情報量が高いほど，理解容易なトピックであることが示された.

6. おわりに

本論文では，LDA によるトピック分析結果の理解容易なトピックを提示するために相互情報量を提案した．相互情報量の大きなトピックを提示することで利用者が理解容易なトピックを提示することが可能になるということが分かった．しかし，今後はどこまでのトピックを利用者に提示するかを検討していく必要がある．また相互情報量を指標とするだけでは相互情報量の低い理解容易なトピックを利用者に提示しないことになってしまう．そこで提示するトピックを絞り込む新しい指標を導入する必要があると考えられる.

さらに今回の実験では理解容易性を 1 人で評価したため，相互情報量と解釈容易性の相関は誰が解釈しても高いのか複数人で評価する必要がある．対象となる論文アブストラクトも複数のキーワードを入力データとして実験する必要がある.

参考文献

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

[2] 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)

[3] 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

[4] 古澤 昂典：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，情報科学技術フォーラム，(2014)

[5] http://gibbslda.sourceforge.net/