討議録に対するトピックモデルの適用

塚井誠人[1]・椎野創介[2]

[1]正会員　広島大学　大学院工学研究院（〒739-8527　東広島市鏡山1-4-1）
E-mail:mtukai@hiroshima-u.ac.jp

[2]学生会員　広島大学　大学院国際協力科（〒739-8528　東広島市鏡山1-5-1）
E-mail:m141298@hiroshima-u.ac.jp

合意形成の質を高めることを目的とした実証分析では，自由記述や討議碌などのテキスト情報の解析から有益な計画情報を抽出する試みが重ねられてきた．しかし従来の手法は，単語の共起情報に基づくクラスタリングによってトピックスを抽出することは可能だったが，1文書が複数のトピックを持つ一般的な場合や，文書内のトピックの推移を明らかにできる手法は，十分に知られていなかった．

本論では，言語情報処理分野で開発が進められてきたトピックモデルの概要と特徴を整理するとともに，トピック抽出による討議分析の可能性について検討する．

Key Words : *Collapsed-Gibbs-Sampling, Latent Dilichllet Allocation, Discouese Management*

1. はじめに

地域計画や都市計画では，計画代替案の最終段階として利害関係者や住民の合意形成による代替案の社会実装が求められる．たとえば自転車通行帯の設置と運用の問題や，バス交通の再編を中心とする地域公共交通網形成計画では，住民や利害関係者を含む検討会や委員会が設置されることが多い．これらの委員会では，地域のローカルな特性を踏まえた実行可能な政策代替案の作成ばかりでなく，実施から一定期間経過した後の政策モニタリングやPDCAサイクルの形成など，長期的な地域の持続可能性の維持などの課題が課されている．

多くのインフラが更新期を迎える我が国では，上述した例にみられるような道路空間の再配分や既存の公共交通インフラの活用を図る政策が多い．それらの政策では，住民がインフラの活用や維持に関して主体的な関与が求められる．そこで，代替案を作成するプロセスの理解や合意形成の質の向上を図るため，討議や情報共有，政策代替案の絞り込みなどの討議をマネジメントする方法の研究が重要と考えられる．羽鳥ら[1]は，公的討議を運営する際の規範論を整理しているが，この分野では規範論も実証分析の蓄積が乏しいことを指摘している．

本論では，討論の実証分析のを目的とした従来のテキストマイニング手法を概観したうえで，機械学習の分野で近年開発されたトピックモデルを適用して，トピック抽出を行う．その際，比較的文章量の少ない討議録データに対して同手法の利点や限界を明らかにする．

2. 既往研究

合意形成に関する実証分析では，膨大な量の文章の特徴を明らかにするため，討議内容の可視化を通じたトピックの要約や計量分析が試みられてきた．しかし実際の討議では様々なトピックが取り上げられる上，関連する語彙は膨大である．このようにテキストデータの統計的な処理は，予めコーディングされた通常のデータよりも極めて大きな困難が伴う．

従来，最もシンプルながら信憑性が高いとされる分析手法は人手によるトピックコードの設定であった．安藤ら[2]は，利害関係者に政策代替案に関する討議を実験的に実施したうえで，収集した討議録から文節ごとのトピックを人手によって整理した．さらにこの研究では，自己組織化マップと主成分分析を段階的に適用して，トピック数を絞り込む手順をとっている．一方塚田ら[3]は，単語の出現頻度や単語間の共起頻度をデータ化して，その傾向から討議主題の推移を明らかにした．具体的にはアンケートの自由記述データに対して，単語間の共起関係をJaccard係数に基づく語のネットワークを形成して可視化した上で，その観察に基づいて結びつきの強い語のグループを抽出して，頻出するトピックを抽出した．

長ら[4]は，インタビュー形式で取得された，より分量の多いテキストデータの分析を試みた．この研究では，発話を単位とするのではなく，文章の意味的な連関の強い部分では同様の単語が繰り返し現れるという特色に着目して，同一のトピックに言及していると思われる分析

単位にテキストデータを分割するText tilingを適用した．さらに以上の手順で作成した分析単位に対してJaccard係数から語のネットワークを形成した．なおこの研究では，トピックの形成に際して，語のネットワークの各ノードごとに，部分ネットワークの指標であるモジュラリティを算出して関連する語群を抽出した．

岩見ら[5]は，語の出現回数と出現の偏りをそれぞれ算出して合成指標化したTF－IDF値に基づくトピックの抽出を行った．この研究では，算出したTF－IDF値から着目する語を絞り込んだ後で，段落を分析単位として着目語の出現を表す多変量データを作成して主成分分析を行った．さらにその負荷量をデータとしてクラスター分析を適用して関連語のグルーピングを行い，得られた単語グループからトピックを抽出した．

佐々木・丸石ら[6]は自己組織化マップを用いて討議のトピックスを抽出する際に，TF－IDF値を用いて抽出する単語の絞り込みを行っている．さらにTF－IDF値を，文書内に出現する全単語について算出したベクトルを討議グループごとに求めて，それらの成す角度によって，討議グループ間の内容の類似性を明らかにした．

長曽我部・榊原[7]は，TF－IDF値によって着目語を絞り込んだ後で，分析単位を発話とするのではなく，発話を前後数回分含む移動平均ウインドウ単位とするアプローチをとっている．この方法では，一定分量の文章があるトピックに言及する特徴を捉えることができる．さらに得られたデータに因子分析を適用してトピックを抽出した上で，その因子負荷量をプロットしてトピックの推移を明らかにしている．

森崎ら[8]は，分析単位を発話としつつも，TF－IDF値を名詞グループと用言グループのペアに適用してトピックを構成する方法を提案している．しかし森崎らの方法ではTF－IDF値を用いても名詞グループと用言グループの組み合わせ数が多くなるという難点を抱えており，最終的に人手による分類に頼らざるを得なかった．

2. トピックモデルの概要

(1) 潜在意味解析[9]

テキストデータから定量データを抽出する手順として，ある文書に現れる語彙の頻度をカウントする方法がある．このようにして得られるデータはBag of words，または文書に現れる語彙のベクトル表現と呼ばれる．

文書 i （ $i=1,...,N$ ）に含まれる語彙 j （ $j=1,...,J$ ）の数をカウントしたベクトルを w_i とする．これらのベクトルを文書単位でスタックした行列 $W=(w_1,...w_i...,w_N)^t$ について，その構造をできるだけ縮約した表現する方法を考えよう．なお t は転置を表す．このとき，W の特異値分解は，以下の式(1)で与えられる．

$$W=U\Sigma V^t \tag{1}$$

特異値分解定理より，行列 Σ は対角行列であり，最大 N 個の相異なる特異値を持つ．また U,V は，それぞれユニタリ行列とその随伴行列である．ここで文書数と語彙数の間には $N<J$ が成り立つ．なお行列 U,Σ,V のサイズは，それぞれ，$N\times N$，$N\times N$，$J\times N$ である．

ここで Σ の対角に現れる特異値を，左上から大きな順に k 個並べた対角行列 $\hat{\Sigma}$ （$k\times k,\ k<N$）と，これから得られる $\hat{W}=U\hat{\Sigma}V^t$ は，W の低ランク近似と呼ばれる．潜在意味解析とは，数学的には W の特異値分解であり，文書数 N よりも少数のトピック数 k によって，元の文書 W を近似する（フロベニウスノルムを最小化する）ことである．ここで，サイズ $N\times k$ の U は文書とトピックの関連を，サイズ $k\times k$ の対角行列 $\hat{\Sigma}$ は，トピックスの強さ（寄与の大きさ）を，サイズ $J\times k$ の V は語彙とトピックの関連を表す．このように潜在意味解析に現れる各行列の要素は対応する成分の関連，すなわち共起の程度を表す．

佐藤・奥村[9]は，潜在意味解析の課題として，1)共起の程度を表す各要素が負の値をとる場合があるため解釈が容易ではないことと，2)抽出されるトピックが特異値分解での性質より互いに直交すること，を挙げている．

一方トピックモデルにおいてこれらの課題は，行列の要素を確率として表現するとともに，必ずしも直交しないトピックスを抽出できる手法が開発されたことによって，方法論的な解決が図られている．

(2) 確率的潜在意味解析とLDA

テキストマイニングが文書から得られる情報縮約を目的としている以上，多くの分析に共通する重要なポイントは，トピックの推定であることがわかる．ただし，トピックはBag of wordsからは直接得られない潜在変数であり，かつ1つの文書は通常，単一のトピックで成立しているわけではなく，多くのトピックスによって構成されている．そこで文書に関してこれらの特徴を考慮できる確率的生成モデルとして，Latent Diriichlet Allocation（LDA）が提案された．

まず文書中の各単語に対応させて，非観測の潜在トピックを表す変数を導入する．文書 d の i 番目の単語を $w_{d,i}$，対応する潜在トピックを $z_{d,i}$ とする．潜在トピックを表すインデックスの集合を K とすると，$z_{d,i}\in(1,2,...,K)$ である．さらに潜在トピック k ごとの語彙 j の分布 $\phi_k\in(1,2,...,K)$ を考えよう．すると

$z_{d,i}$ が指定する潜在トピック k の下で，単語 $w_{d,i}$ が分布 ϕ_k に従って出現するという構造が考えられる．さらに以下の展開のため，ここで文書 d ごとのトピック k の分布 $\theta_d \in (1,2,...,N)$ も導入しておく．

LDAでは離散値の出現確率に関する分布を指定することによって，それらの出現確率を算出する．具体的には，多項分布とDirichlet分布が用いられる．多項分布は，各試行においてK種類の値をとる確率変数の生起確率 $\boldsymbol{\pi}$ が，

$$\boldsymbol{\pi} = (\pi_1,..,\pi_K) \qquad \left(\sum_i \pi_i = 1\right) \tag{2}$$

のとき，独立な試行列 $\mathbf{x} = (x_1,..,x_n)$ の生成確率は，$x_i = k$ となる回数を n_k であらわすと，

$$p(\mathbf{x}|\boldsymbol{\pi}) = \prod_{i=1}^{n} p(x_i|\boldsymbol{\pi}) = \prod_i \pi_i^{n_k} \tag{3}$$

となる．以上より各事象kの生起回数：$p(\{n_k\}_{k=1}^K|\boldsymbol{\pi}) = (n_1,..,n_K)$ の分布は，

$$p(\{n_k\}_{k=1}^K|\boldsymbol{\pi}) = \frac{n!}{\prod_{k=1}^K n_k!}\prod_{k=1}^K \pi_k^{n_k} \tag{4}$$

で定義される多項分布 $\mathrm{Multi}(\{n_k\}_{k=1}^K|\boldsymbol{\pi})$ に従う．なお，式(4)から明らかなように，各試行における x_i は，$n=1$ の多項分布に従うことがわかる．

多項分布のパラメータとなっている事象kの生起確率 $\boldsymbol{\pi}$ の確率分布を考える．ここで $\boldsymbol{\pi}$ は座標の和が1で定義される空間であり，単体と呼ばれる．Dirichlet分布はこの単体上で定義される確率分布であり，ハイパーパラメータ $\boldsymbol{\alpha} = (\alpha_1,..,\alpha_k)$ を用いて，

$$p(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)}\prod_{k=1}^K \pi_i^{\alpha_k - 1} \tag{5}$$

と表わされる．ここで $\Gamma(x)$ は，ガンマ関数である．

LDAは，語彙とトピックの出現確率がそれぞれ多項分布に従い，それぞれの出現確率がそれぞれのハイパーパラメータを有するDirichlet分布に従う構造を仮定した，Bag of wordsの確率的生成モデルである．確率変数と分布の関係は，次のように整理できる．

$$\begin{aligned} \theta_d &\sim \mathrm{Dir}(\boldsymbol{\alpha}),\ d = 1,...,N \\ \phi_k &\sim \mathrm{Dir}(\boldsymbol{\beta}),\ k = 1,...,J \end{aligned} \tag{6}$$

ここでハイパーパラメータ $\boldsymbol{\alpha}$，$\boldsymbol{\beta}$ はそれぞれ，文書数 N，語彙数 J の次元をもつ．さらに，θ_d，の下で ϕ_k

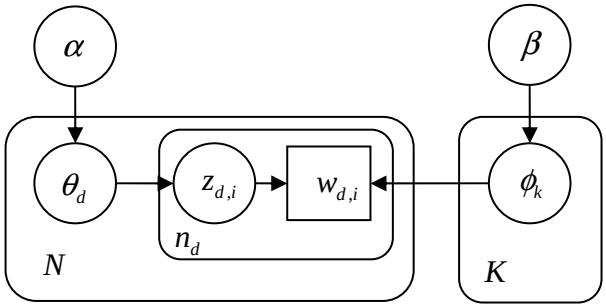

図1　LDAのグラフィカルモデル

潜在トピックと単語は，以下のように生成される．

$$\begin{aligned} z_{d,i} &\sim \mathrm{Multi}(\theta_d),\ i = 1,...,n_d \\ w_{d,i} &\sim \mathrm{Multi}(\phi_{z\_d,i}),\ i = 1,...,n_d \end{aligned} \tag{7}$$

なお n_d は，文書 d の総単語長を表す．LDAのグラフィカルモデルを図1に示す．同図が示すように，LDAの全確率の結合分布は，次のように構造化されている．

$$p(\mathbf{w},\mathbf{z},\boldsymbol{\theta},\varphi|\boldsymbol{\alpha},\boldsymbol{\beta}) = p(\mathbf{w}|\mathbf{z},\boldsymbol{\theta})p(\mathbf{z}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\alpha})p(\boldsymbol{\varphi}|\boldsymbol{\beta}) \tag{8}$$

ここで，

$$p(\mathbf{w}|\mathbf{z},\boldsymbol{\theta})p(\mathbf{z}|\boldsymbol{\theta}) = \prod p(w_{d,i}|\phi_{z\_d,j})p(z_{d,i}|\theta_d)$$

$$p(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \prod_d p(\theta_d|\boldsymbol{\alpha})$$

$$p(\boldsymbol{\varphi}|\boldsymbol{\beta}) = \prod_k p(\phi_k|\beta)$$

である．

LDAによって確率的潜在意味解析を行うモデルは，トピックモデルと呼ばれる[10]．

(3)　トピックモデルの利点と限界

トピックモデルは，言うまでもなく潜在変数であるトピックkが推定できるという利点を有する．しかし機械学習の観点からは，構造が柔軟であるという利点が挙げられる．つまり $\theta_{d,k} = p(k|d)$，および $\phi_{k,j} = p(j|k)$ と表記すれば，文書dで語彙jが出現する確率 $p(j|d)$ は，

$$p(j|d) = \sum_{k=1}^{K} p(j|k)\,p(k|d) \tag{9}$$

と表すことができる．式(9)が示すように，トピックkを介して語彙の出現確率を算出しているので，実際にはその部分に現れない語彙であっても，出現確率が算出できる．また文書とトピックの関係を表す ϕ_d では1つの文書が確率的（または潜在的）に複数のトピックを有する

ことが表現されているので，文書ごとのトピックの出現分布の違いが表現できる．またトピックと語彙の関係を表すϕ_k を観察することにより，そのトピックが表わす内容を推定できる．

またトピックモデルでは同一の語彙が別のトピックの下で出現できる構造になっている．これは，各語彙の観点からは語彙が文脈に依存して意味が異なるという語義の多義性の問題を，集合d （つまりトピック）を用いることによって解消している，とも理解できる．

ハイパーパラメータの推定に関しては，Dirichlet分布が多項分布の共役事前分布となっているため，Gibbs-samplingによる統計的学習を行うことができる．これらのアルゴリズムに関しても多くの研究例が蓄積されており，ある程度の文章量の文書が多数得られる場合，ハイパーパラメータが安定的に推定できることが示されている[10]．

討議録を対象とする本研究では，比較的文章量の少ない文書を対象にトピックを抽出しなければならない．たとえば討議録の発話単位の文章量は極めて短い場合がある．文書全体の単語数が有限な以上，分析単位に含まれる単語数と文書数（分析単位）はトレードオフの関係にあり，適切の規範の下で最適な設定を見いだせる可能性がある．なおこの問題は，学習の効率性と得られたトピックの解釈可能性の両面から検討する余地がある．また学習効率の向上に関連して，入力情報からの低頻度語や付属語などの除去の問題がある．これらはストップワードと呼ばれる語であり，適切な処理が必要となる．この問題はテキスト解析に共通する課題だが，文書量が少ない討議録の分析では重要度が高いと思われる．

討議のマネジメントの観点からは，トピックの抽出にとどまらず，トピックや発話（これまでの議論では文書）が討議の流れに及ぼす影響を明らかにする分析が求められる．たとえば発話やトピックの重要性など，テキストに関連する付帯情報が得られる場合，これらのラベル情報とトピックの関連性を明らかにする分析が求められる．これは機械学習の分野では，教師あり学習問題と呼ばれる．Flahertyら[11]はトピックモデルの拡張として，トピックとラベルの関連性を明らかにできる結合トピックモデルを提案している．一方で発話とラベルの関連性に関しては，筆者の知る限りでは，分析手法の開発が待たれる状況にある．

3. トピックモデルの適用例

計算例は，発表時に示す．

参考文献

1) 羽鳥剛史，小林潔司，鄭蝦榮：討議理論と公的討論の規範的評価，土木学会論文集 D3，69-2, 101-120, 2013
2) 塚田慎也，森田哲夫，西尾敏和，湯沢昭：自由記述データに着目した限界自治体における生活質評価に関する分析，日本建築学会計画系論文集，80-708，361-368，2015
3) 安藤章，森川高行，三輪富生，山本俊行：フォーカスグループインタビューの討議分析からみた市民のＰＤＳに対する評価特性，第 40 回土木計画学研究・講演集，Vol.40，CD-ROM，2009.
4) 長尚希，室町泰徳，板谷和也：計量的言語処理を利用した大規模交通プロジェクトに関する経験知識の抽出に関する研究，都市計画論文集 47(3), 793-798, 2012
5) 岩見麻子，大野智彦，木村道徳，井手慎司：公共事業計画策定過程の議事録に対するテキストマイニングによる議論内容の把握に関する基礎的研究，土木学会論文集 G，68-6，411-418，2012
6) 佐々木邦明，丸石浩一：テキストマイニングを用いたワークショップの討議内容の特徴把握と可視化に関する研究，都市計画論文集，46(3), 1039-1044, 2011
7) 長曽我部まどか，榊原弘之：ワークショップにおける相互補完的対話の分析，都市計画論文集，50(1), 28-36, 2015
8) 森崎孔太，塚井誠人，難波雄二，桑野将司：司会者の関与が討議参加者の納得に及ぼす影響，土木学会論文 D3，70-1, 28-43, 2014.
9) 佐藤一誠，奥村学：トピックモデルによる統計的潜在意味解析，コロナ社（自然言語処理シリーズ），2015
10) 岩田具治：トピックモデル，講談社（機械学習プロフェッショナルシリーズ），2015
11) Flaherty P, Giaever G, Kumm J, Jordan MI, Arkin AP.: A latent variable model for chemogenomic profiling, Bioinformatics, 21(15):3286-93, 2005.

（2015.07.31受付）

An application of “topic model” to discussion records

Makoto TSUKAI, Sosuke SHIINO