# ニューラルネットワークを用いたトピック遷移モデリングに関する検討

内田 脩斗　　吉川 大弘　　古橋 武
名古屋大学大学院 工学研究科
{uchida, yoshikawa, furuhashi}@cmplx.cse.nagoya-u.ac.jp

## 1 はじめに

近年，Twitter をはじめとする SNS（Social Networking Service）の普及に伴い，膨大なテキストデータから有用な知識を獲得するテキストマイニングの取り組みが活発化している [1][2]．なかでも，テキストデータの時系列性に注目することで，株価 [3] や視聴率 [4] など，社会現象の予測に用いた研究が多く報告されている．しかし，ユーザごとのテキストデータの系列性に注目した研究は少なく，新たなデータ分析の視点となることが期待できる．そこで本稿では，テキストデータに基づくユーザの行動モデリングについて検討する．つまり，ある投稿を発信したユーザが次にどのような投稿を行うか，という投稿間の遷移パターンのモデル化である．

本稿では，文書表現として LDA を，また，遷移モデリングにニューラルネットワークを用いた手法を提案する．さらに，学習済みモデルを利用した遷移パターンの分析法を提案し，実際に分析を行った結果を示す．本手法により，投稿間の遷移関係の分析を容易にし，ある投稿が発信された原因・根拠の明確化につながることが期待できる．

## 2 関連研究

行動モデリングとは，ユーザの行動の依存関係をモデル化することであり，行動の予測やサポートに応用されている．なかでも，マルコフ連鎖における状態遷移行列は代表的なモデリング手法の一種である．[5] では，音楽視聴ログを用いて，視聴パターンのモデリングを行い，その遷移分析結果を報告している．また，[6] では，ニューラルネットワークに事前確率に基づくバイアスを加えることで，柔軟な状態遷移を表現可能であることを報告している．しかし，テキストデータを用いた行動モデリングに関する研究は少なく，新たな研究対象となり得る．

また，膨大な文書データの解析において，注目されている技術の一つとしてトピックモデルがある．トピックモデルとは，文書中に含まれる単語の生成過程を確率的にモデリングすることで，文書に潜在しているトピックを抽出する手法である．代表的なものに，Latent Dirichlet Allocation(LDA)[7] がある．LDA は，その拡張性の高さが広く知られており，言語分野だけでなく，音声認識や画像処理など多くの分野に適用されている．[8] では，文書の時系列性を考慮したトピックモデルが提案されている．また，[9] では，音楽視聴ログを用いて，時間変化するユーザの興味およびアイテムの追跡を可能としたトピックモデルを提案している．しかし，これらの手法はあるトピックの衰勢を追跡するモデルであり，各ユーザの投稿を追跡する行動モデリングとは異なる．

最も本研究に関連している手法として，TM-LDA[10] では，Twitter におけるツイートの前後関係を表現する状態遷移行列を導出し，トピックの遷移パターン分析を示している．本手法は，社会現象などの発生原因や根拠を提示することが容易な分析モデルといえる．

## 3 従来手法と提案手法

### 3.1 Temporal-LDA

TM-LDA[10] では，LDA により生成されるトピック分布を用い，マルコフ性を仮定することで，文書の前後関係を表現する状態遷移行列 $T$ を式 (1) により獲得する．

$$T = (A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}B \tag{1}$$

このとき，$A$ は過去の投稿データをトピックベクトル化した行列，$B$ は未来の投稿データをトピックベクトル化した行列を表している (図 1)．

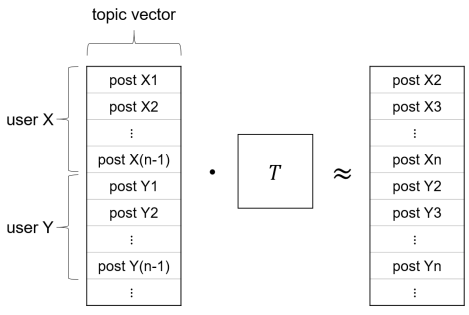



**図 1:** TM-LDA の構成

ただし，この手法の改善点として，大きく 2 つのことが考えられる．まず，投稿間の関係を表現している $T$ の表現能力が限定的であることが挙げられる．$T$ は，$(K \times K)$ の行列 ($K$：トピック数) であり，これを拡張することで予測性能の向上が期待できる．次に，入力データ形式が文書のトピックベクトル以外受け付けないことが挙げられる．$A$ は前提として，各行ベクトルが文書のトピックベクトル，つまり，

$$\sum_{d=1}^{K} w_d = 1 \quad (0 \leq w_d \leq 1) \tag{2}$$

という制約の上 ($B$ も同様) で，式 (1) が成り立っている．そのため，データに付随する属性情報や時間情報といったメタ情報を加えることなどが困難である．ユーザの性別や年齢により，投稿の内容に差異が生まれることは容易に想像がつき，また，予測モデルの特徴量として加えることで予測性能の向上が期待できる．さらに，遷移分析においても，属性情報を加えたより詳細な分析が可能となると考えられる．ただし，本稿では，メタ情報の付加方法については割愛する．

### 3.2 Malkov Chain Neural Network - LDA

MCNN-LDA では，状態遷移行列 $T$ に対応する部分に，ニューラルネットワークを適用する (図 2)．なお本手法では，TM-LDA と同様に，マルコフ性を仮定している．これにより，3.1 で問題点として挙げた，モデルの限定的な表現能力と拡張性の乏しさを克服することが可能であると考えられる．前者（モデルの限定的な表現能力）は，隠れ層を導入することで容易に表現幅を広げることが可能である．また，後者は，ニューラルネットワークの入力データ形式に制約がないため，容易に入力次元を拡張することができ，新たな特徴量を加えることが可能である．

ただし，制約として，ニューラルネットワークの出力はトピックベクトルである必要がある．つまり，式 (2) を満たす必要がある．そのため，出力層には以下に示す $softmax$ 関数 を導入することで，予測値を確率分布として扱えるようにする．

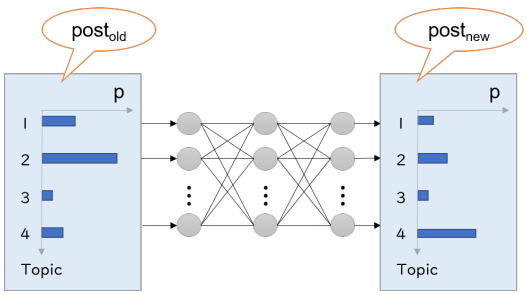



**図 2:** MCNN-LDA の構成

$$softmax(x_i) = \frac{exp(x_i)}{\sum_{d=1}^{K} exp(x_d)} \tag{3}$$

さらに，回帰予測である点を考慮し，損失関数には MSE (Mean Square Error) を用いる．

## 4 実験

### 4.1 実験方法

あるユーザの投稿群を，古い投稿から 1, 2,..., $n$ とする．

1. 1〜$(n-1)$ の投稿を学習データとし，LDA を用いてトピック分布を生成する．
2. 学習データをトピックベクトル化する．
3. 投稿の前後関係から，学習用のペアデータを生成する (図 1)．
4. モデル（TM-LDA(式 (1))/MCNN-LDA(図 2)）の学習を行う．ただし本実験では，MCNN-LDA の入力には，2. で得られたトピックベクトルのみを用い，メタ情報は用いない．
5. $(n-1)$ の投稿をモデルの入力として，次投稿のトピック分布を予測し，$n$ の投稿との誤差を評価する．

本実験では，評価指標として $perplexity$(式 (4)) を用いた．

$$perplexity = exp(\frac{\sum_{d=1}^{M}\sum_{n=1}^{N} logp(\mathbf{w}_{dn})}{\sum_{d=1}^{M} N_d}) \tag{4}$$

ここで，$M$ は文書数，$N_d$ は文書 $d$ の単語数，$p(\mathbf{w}_{dn})$ はある文書 $d$ 内の $n$ 番目の単語が生成される確率を表している．$perplexity$ は，文書に出現する単語の選択肢の数を表しており，値が小さいほど予測精度が高いことを意味する．

### 4.2 FKC コーパス

本実験では，不満買取センターにて収集されている FKC コーパス [11] を用いた．これは，ユーザの感じた不満を自由記述形式で収集するサービスで，データ数の多さ (表 1) や，年齢や性別などのメタデータの豊富さが特徴である．本実験では，投稿数が 3〜40 のユーザを対象とした．また，前処理として投稿データには，ストップワードの除去と名詞の抽出を行っている．なお表 1 において，投稿データ数に比べて学習・テストに用いているデータ数が少ないのは，投稿数 2 以下や 41 以上のユーザを除いたためであり，多いものでは 1 万件以上の投稿をしているユーザも存在しているためである．また，本実験では，各ユーザの最終

投稿を予測する形になるため，学習ユーザ数 = テストデータ数 となる．

**表 1:** FKC コーパスのデータ情報

| 対象期間 | 2015/03/18 - 2017/03/12 |
|---|---|
| 投稿データ数 | 5,248,820 |
| ユーザ数 | 106,173 |
| 学習データ数 | 544,922 |
| 学習ユーザ数 | 48,703 |
| テストデータ数 | 48,703 |

### 4.3 ネットワーク構成

本実験で設計したニューラルネットワークのパラメータを表 2 に示す．本実験では，隠れ層が 1 層のネットワークを用いた．なお，$K$ はトピック数である．

**表 2:** MCNN-LDA のパラメータ

| パラメータ | 値 |
|---|---|
| 層数&次元数 | K-K-K |
| Optimizer | Adam |
| 初期学習率 | 0.001 |
| 活性化関数 | relu |
| 損失関数 | MSE |
| epochs | 40 |
| ミニバッチサイズ | 64 |
| 終了条件 | early stopping |

### 4.4 実験結果

表 3 に，実験結果を示す．これより，提案手法では *perplexity* が従来手法より低下しており，予測精度の向上が確認された．予測モデルの設計にニューラルネットワークを用いたことで，より細かな確率予測が可能となったと考えられる．

## 5 トピック遷移分析

本章では，4 章で得られた予測モデルを利用したトピック遷移分析法について言及する．

### 5.1 状態遷移行列 $T$ の生成

TM-LDA では，式 (1) により生成される $T$ を直接分析することで，特徴的なトピック遷移を可視化することができる (図 3(a))．縦軸が遷移前のトピック，横軸が遷移後のトピックを表しており，確率の大きい成分ほど特徴的なトピック遷移を表している．一方，MCNN-LDA では，予測モデル自体が複雑な構造をしているため，直接的に $T$ を可視化することができない．そこで，モデルの入力データとして，単位行列 $I$ を考える．これにより，あるトピックから各トピックへの遷移確率が予測でき，予測結果を擬似的な状態遷移行列 $T$ として扱うことが可能となる．以上の方法を用いて，生成された状態遷移行列 $T$ を図 3(b) に示す．ただし，このままでは，特徴的な遷移を分析しづらいため，次節で閾値除去を行う．

**表 3:** 予測性能評価結果

| トピック数 | TM-LDA | MCNN-LDA |
|---|---|---|
| 50 | 1840.1 | 1834.9 |
| 100 | 2145.0 | 2136.7 |
| 150 | 2469.1 | 2441.7 |
| 200 | 2822.7 | 2754.8 |

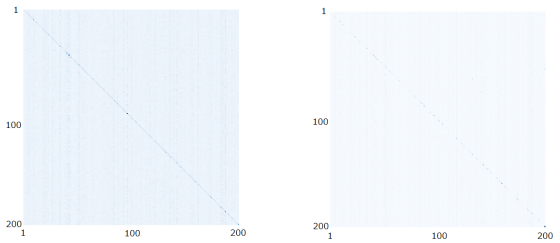
### 5.2 特徴的トピック遷移の抽出

#### 5.2.1 閾値の決定法

図 3(a)，3(b) において，対角成分の重みが大きいことから，これらを基準に閾値を決定する．対角成分の平均を $\bar{t}$，対角成分以外の標準偏差を $\sigma$ とし，式 (2) を用いて閾値を算出する [10]．

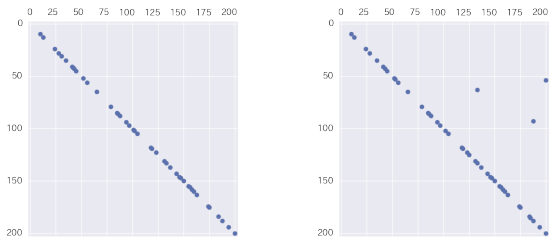
$$threshold = \bar{t} + 5 \times \sigma \tag{5}$$

実際に閾値除去を行った結果を図 4(a)，4(b) に示す．ただし，$threshold_{TM} = 0.044, threshold_{MCNN} = 0.037$ のように，閾値が手法により異なるため，今回は下限の 0.037 に揃えている．図 4(b) より，MCNN-LDA では，対角成分以外にも大きな重みを分散させていることがわかる． このことから，ニューラルネッ




**(a)** TM-LDA **(b)** MCNN-LDA

**図 3:** 状態遷移行列 $T$ の可視化




**(a)** TM-LDA **(b)** MCNN-LDA

**図 4:** 特徴的トピック遷移の可視化

トワークを用いたことで，より細かな重み配分が可能となり，4.4 で示した性能向上につながっている要因の一つとなっていると考えられる．

### 5.2.2 特徴的トピック遷移

表 4 に，図 4(b) の対角成分以外の特徴的トピック遷移を示す．かっこ内の数字は，トピック番号を表しており，各トピックには，トピックの内容を包括的に表すラベルを著者らが主観的に付与した．

本実験では，不満データを包括的に利用したため，表 4 の結果は，ユーザ属性と関連性が高いことが考えられる．実際，図 5 のように，ユーザが職業別に偏って分布しており，専業主婦に特徴的なトピックや，会社員に特徴的なトピックが表 4 に現れていることが確認できる．また，遷移に注目すると,「ネイル」から「美容院」への遷移が特徴的であり，どちらも美容に関するトピックであるが，逆方向の遷移は特徴的トピック遷移として抽出されていないことが興味深い．つまり，そこには何らかのユーザの行動パターンや想起パターンがある可能性が考えられる．これを元に更なる分析を進めることで，有用な知見の獲得につながることが期待できる．

## 6 まとめと今後の課題

本稿では，テキストデータにおけるユーザの行動モデリングとして，ニューラルネットワークを用いた投稿トピックの遷移モデリング手法を提案した．また，従来の TM-LDA との性能比較実験を行い，予測性能の向上を確認した．

今後の課題として，3.2 で取り上げたメタ情報の付与や属性情報を利用した詳細な遷移分析が挙げられる．

## 謝辞

本研究では，株式会社 Insight Tech が国立情報学研究所の協力により研究目的で提供している「不満調査データセット」を利用した．

**表 4:** 特徴的なトピック遷移

| 遷移前 | 遷移後 | 遷移確率 |
|---|---|---|
| 株式会社 (54) | ゲーム (200) | 0.073 |
| ネイル (63) | 美容院 (133) | 0.063 |
| 契約 (93) | 仕事環境 (188) | 0.039 |

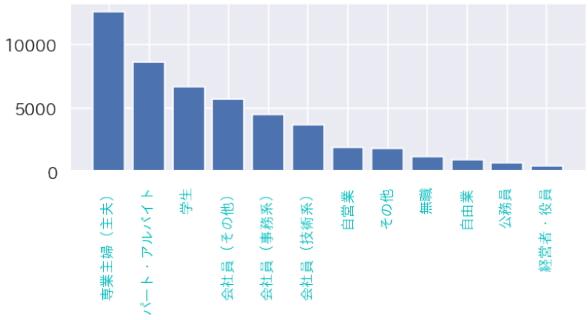



**図 5:** 職業別ユーザ数

## 参考文献

[1] P.P Khaing and N. New. Adaptive methods for efficient burst and correlative burst detection. *In: 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*, 2017.

[2] T. Ma, Y. Zhao, H. Zhou, Y. Tian, A. Al-Dhelaan, and M. Al-Rodhaan. Natural disaster topic extraction in sina microblogging based on graph analysis. *Expert Systems with Applications*, Vol. 115, pp. 346–355, 2019.

[3] C.S. HO, P. Damien, B. Gu, and P. Konana. The time-varying nature of social media sentiments in modeling stock returns. *Decision Support Systems*, Vol. 101, pp. 69–81, 2017.

[4] A. Crisci, V. Grasso, P. Nesi, G. Pantaleo, I. Paoli, and I. Zaza. Predicting tv programme audience by using twitter based metrics. *In Multimedi a Tools and Applications*, Vol. 77, pp. 12203–12232, 2018.

[5] F. Figueiredo, B. Ribeiro, J.M. Almeida, and C. Faloutsos. Tribeflow: Mining & predicting user trajectories. *In: International Conference on World Wide Web*, pp. 695–706, 2016.

[6] M. Awiszus and B. Rosenhahn. Markov chain neural networks. *arXiv preprint arXiv:1805.00784*, 2018.

[7] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.

[8] D.M. Blei and J.D. Lafferty. Dynamic topic models. *Proceedings of the 23rd international conference on Machine learning*, pp. 113–120, 2006.

[9] T. Iwata, T. Yamada, Y. Sakurai, and N. Ueda. Sequential modeling of topic dynamics with multiple timescales. *ACM Trans. Knowledge Discovery from Data (TKDD)*, Vol. 5, No. 4, pp. 1–19, 2012.

[10] Y. Wang, E. Agichtein, and M. Benzi. Tm-lda: efficient online modeling of latent topic transitions in social media. *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 123–131, 2012.

[11] K. Mitsuzawa, M. Tauchi, M. Domoulin, M. Nakashima, and T. Mizumoto. Fkc corpus: a japanese corpus from new opinion survey service. *In proceedings of the Novel Incentives for Collecting Data and Annotation from People: types, implementation, tasking requirements, workflow and results*, pp. 11–18, 2016.