The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

FKC コーパスを用いた トピック遷移分析手法に関する検討

A Study on Topic Transition Analysis Method Using FKC Corpus

内田脩斗 *1 Shuto Uchida
吉川大弘 *1 Tomohiro Yoshikawa
古橋武 *1 Takeshi Furuhashi

*1名古屋大学大学院工学研究科
Graduate School of Engineering Nagoya University

With the spread of SNS, text mining to acquire useful knowledge from enormous data has been active. Moreover, much attention has been paid to time series data and many studies predicting social phenomena have been reported. However, there are few studies focusing on the seriesiness of text data for each user and it can be expected to be a new viewpoint of data analysis in near future. In this paper, we discuss the behavior modeling of users based on text data. We propose a method using LDA as the document representation and Neural Network as the transition modeling. By using Neural Network, it becomes possible to expand the expression ability and flexibility, and then, the improvement of modeling can be expected compared with the conventional method. In fact, we carry out an experiment and report that the performance of the proposed method is improved. We also show an example of topic transition analysis and mention the practicality of the method.

1. はじめに

近年，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 [Khaing 17]．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 [HO 17] や視聴率 [Crisci 18] 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．また，ユーザの日常生活における行動パターンや Web 閲覧行動など，行動モデリングに関する研究も広くなされている [Shen 07]，[本村 09]．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数少なく，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．具体的には，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を試みる．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．ニューラルネットワークを用いることで，モデルの表現能力の拡大と柔軟性を取り入れることが可能となる．また本稿では，FKC コーパス [Mitsuzawa 16] を対象とし，従来手法との遷移モデリング性能の比較を行う．FKC コーパスは，ユーザの感じた不満を収集したコーパスであり，ユーザの生活環境や行動履歴が色濃く反映されている．つまり，ユーザの投稿間には，ユーザ特有の行動パターンに基づくトピックの遷移が想定され，分析価値の高いデータであると考えられる．さらに，提案手法を用いたトピック遷移分析の一例を示す．FKC コーパスにおけるトピック遷移分析では，不満投稿間の因果関係（行動パターンや想起パターン）を捉えることが可能であり，ある社会現象などの発生原因・根拠のより深い考察を可能とする分析ツールとな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，日常生活行動や購買行動の予測やサポートに応用されている [本村 09]，[Shen 07]．なかでも，次状態が現在の状態にのみ依存するというマルコフ連鎖の概念を取り入れた手法が数多く報告されている [Gambs 12]，[小澤 13]，[Blanchet 13]．[Figueiredo 16] 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示している．また，[Awiszus 18] 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，新たなデータ分析の視点となることが期待できる．

また，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）[Blei 03] 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，画像処理や音声認識など多くの分野に適用されている [Wang 09]，[Spina 15]．[Blei 06] 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おり，あるトピックの単語分布がどのように変化していくか追跡が可能なモデルとなっている．また [Iwata 12] 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．

最も本稿に関連している手法として，TM-LDA[Wang 12] 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数式を用いて導出し，トピックの遷移パターン分析を示している．本手法は，テキストデータにおいて行動モデリングを試みた手法であり，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[Wang 12] 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

連絡先: 内田脩斗，名古屋大学大学院工学研究科，名古屋市千種区不老町，TEL:052-789-2793，FAX:052-789-3166，uchida@cmplx.cse.nagoya-u.ac.jp

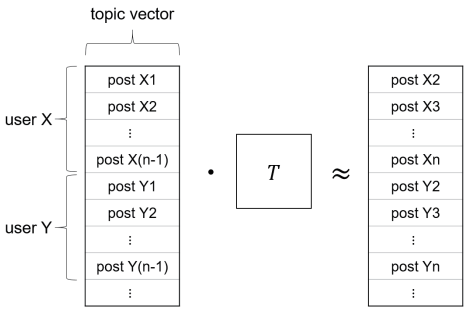


図 1: TM-LDA の構成

を表している（図 1）．

ただし，この手法の改善可能な点として，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データ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．

3.2 Markov 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，3.1se 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（図 3）．

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 $softmax$ 関数（式 (3)）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．

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

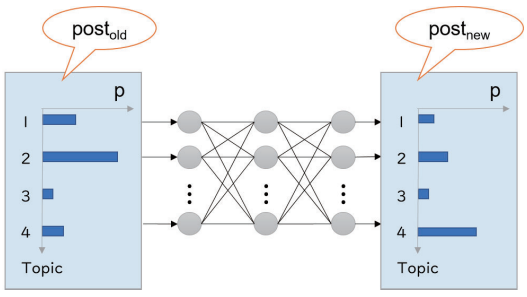


図 2: MCNN-LDA の構成

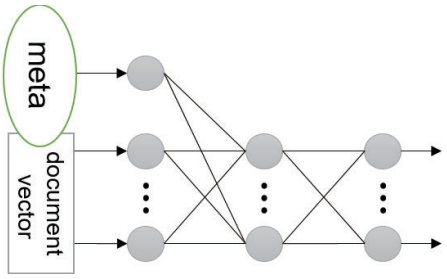


図 3: メタデータを用いた場合の MCNN-LDA

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．

次に，投稿データに付随しているユーザのメタデータ（居住地や性別，年齢など）を，ニューラルネットワークに付加する手法について示す．まず，メタデータが連続値（年齢や投稿時間など）の場合，図 3 のようにネットワークの入力次元を一つ拡張することで，新たな予測モデルの特徴量として利用することが可能となる．また，離散値（居住地や性別など）の場合，対応した要素のみを 1 で表現する one-hot ベクトルを用いることで，ネットワークの特徴量として加えることができる．つまり，メタデータが居住地であれば 47 次元，性別であれば 2 次元の入力次元拡張となる．

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（1 と 2, 2 と 3 ,..., $(n-2)$ と $(n-1)$）を生成する（図 1）．
4. 予測モデルの学習を行う（TM-LDA（式 (1)），MCNN-LDA（図 2 or 図 3））．
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．

なお本実験では，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 [Mitsuzawa 16] 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，豊富な投稿データやメタデータ（居住地，性別など）を用いることができる．また，個々の不満投稿には，ユーザの生活環境や行動履歴が反映されていると考えられ，その投稿間には，ユーザ特有の行動パターンに基づくトピックの遷移が現れることが期待できる．

表 1 に，実験で用いた FKC コーパスのデータ詳細を示す．本実験では，投稿数が 3〜1000 のユーザを対象とした．また，居住地・職業・性別の属性に欠損のないユーザを抽出した．さらに，前処理として，投稿データには，ストップワードの除去

表 1: FKC コーパスのデータ詳細

対象期間	2015/03/18 - 2017/03/12
投稿データ数	3,094,175
ユーザ数	64,196
学習ペア数	2,965,783
テストペア数	64,196
ボキャブラリ数	87,780

と名詞の抽出，低頻度語の除去を行った．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，表 1 のように，ユーザ数 = テストペア数 となる．

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．また，メタデータ（次元数：$meta$）を特徴量として用いた場合，入力層の次元数は $K + meta$ となる．

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	30
ミニバッチサイズ	64
終了条件	early stopping

4.4 実験結果

図 4 に，各トピック数におけるモデルの予測精度を示す．LDA のランダム性を考慮し，5 試行の平均と標準偏差を示している．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．また，トピック数が増加するとともに，TM-LDA と MCNN-LDA の差が広がっていることがわかる．トピック数が大きいことは，文書表現が豊かであることを意味しており，次投稿予測においては，より難しいタスクとなる．MCNN-LDA では，トピック数の増加に対しても，$perplexity$ がやや低下していることから柔軟に予測モデルを構築できることを表していると考えられる．

次に，メタデータを付与した場合の予測精度（K : 200）を図 5 に示す．なお，かっこ内の数字は，メタデータの次元数を表している．この結果より，メタデータを付与することで，単純な MCNN-LDA よりも予測精度が向上していることがわかる．これにより，予測モデルにメタデータを取り入れることの重要性が確認された．ただし，MCNN+居住地 (47) については，検討の余地を残す結果となった．これは，地方単位にすることや, 都市とそれ以外の 2 分類にするなど，特徴量にまとまりをもたせることで，予測精度の改善が期待される．

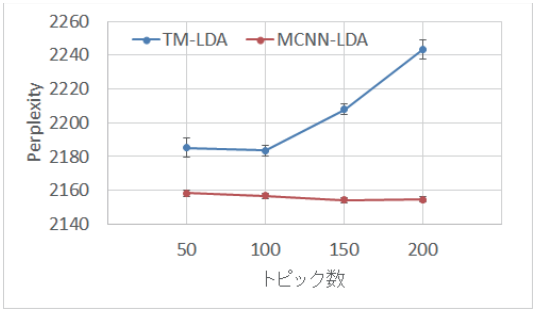


図 4: 各トピック数における予測精度の比較

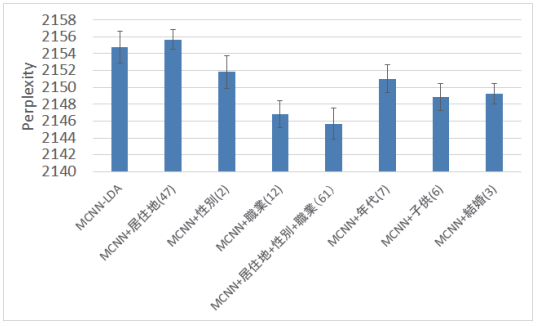
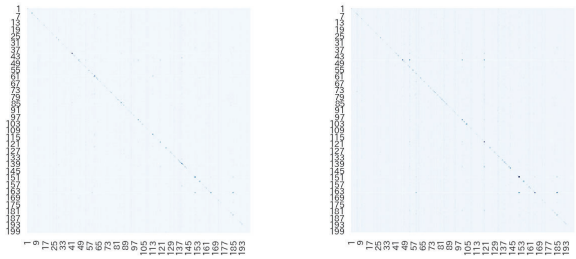


図 5: メタデータ付与による予測精度の比較 (K : 200)

5. トピック遷移分析

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 を用いることで，特徴的なトピック遷移を容易に分析することができる．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 を用いることができない．そこで，MCNN-LDA 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．また，メタデータを用いた MCNN-LDA の場合，それぞれの特徴量に対応した状態遷移行列 T が生成される．以上の方法を用いて，生成された状態遷移行列 T を図 6(a)，(b) に示す．なおここでは，メタデータとして職業を用いた場合のモデルを用いている．



(a) 学生　　(b) パート・アルバイト

図 6: 状態遷移行列 T の可視化（職業別）

ただし，このままでは扱いにくいため，次節において閾値除去による特徴的トピック遷移の抽出を行う．

5.2 閾値除去による特徴的トピック遷移の抽出

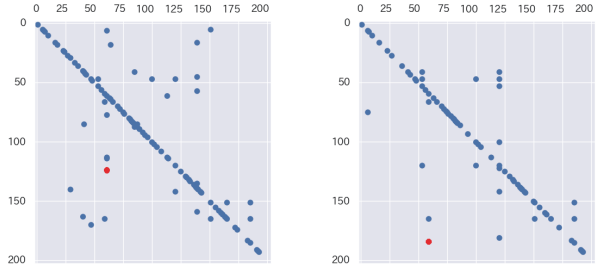
状態遷移行列 T の分析を容易にする手法として，閾値除去による特徴的トピック遷移の抽出法 [Wang 12] を示す．ここでは，対角成分の平均を $\bar{t}$，対角成分以外の標準偏差を σ とし，式 (5) を用いて閾値を算出する．これにより，実際に閾値除去を行った結果を図 7(a)，(b) に示す．

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

また，LDA を用いて生成されたトピックの一例を表 3 に示す．かっこ内の数字は，トピック番号を表しており，各トピック名は，著者らが主観的に付与した．また，トピック番号は図 7 の番号と対応している．行番号が遷移元トピック番号，列番号が遷移先トピック番号を表している．これより，図 7(a) では，60 列目（天気）と 138 列目（広告），図 7(b) では，58 列目（店）と 119 列目（家電）がそれぞれ特徴的なトピック遷

表 3: トピックの例（トピック名とトピック所属確率上位 5 単語を表示）

Topic(number)	買い物（123）	天気（60）	広告（138）	人（183）	店（58）	家電（119）
5 Top Words	買い物	物	表示	客	コンビニ	家電
	高い	洗濯	広告	従業員	スーパーマーケット	エアコン
	見た目	天気	邪魔	アルバイト	スーパー	メニュー
	一般	気候	自動	耳	セブンイレブン	音
	買い	部屋	動画	ホット	店	掃除機



(a) 学生　　**(b)** パート・アルバイト

図 7: 特徴的トピック遷移の可視化（職業別）

移先であることがわかる．これらは，ユーザの職業と関連性の高いトピックであると考えられる．また，トピック遷移の一例を示すと，図 7(a) では，123（買い物）→ 60（天気）のトピック遷移，図 7(b) では，183（人）→ 58（店）のトピック遷移が特徴的であることがわかる．これらは，それぞれが関連性のあるトピックであり，かつ，単方向性が表れている（双方向のトピック遷移の場合，対称点にプロットされる）と考えられる．また，これらは，各々の職業におけるユーザ特有の行動パターンや投稿パターンを表現しているものであり，ある投稿が発信された原因や根拠づけを解釈する際の糸口としての利用が期待できる．さらに，あるトピックを解釈する際に，遷移前や遷移後のトピックを補助情報として用いることが可能であり，トピック解釈性の向上にも役立つと考えられる．

6. まとめ

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トピック遷移モデリング手法を提案した．また，従来手法である TM-LDA との性能比較実験を行い，提案手法による予測精度の向上を確認した．さらに，提案手法を用いたトピック遷移分析により，ユーザの属性ごとにトピック遷移に異なる特徴が現れることが確認でき，より詳細な分析への糸口となり得ることを示唆した．

今後の課題として，さらなるトピック遷移予測精度の向上が挙げられる．これは，マルコフ性の解除やネットワークの拡張などの工夫が有効であると思われる．また，遷移分析のより詳細な解析法が求められる．現状では，名詞に限定した分析を行っているが，さらに，形容詞やかかり受け情報を用いることで，より明確な分析が可能となることが期待される．

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．

参考文献

[Awiszus 18] Awiszus, M. and Rosenhahn, B.: Markov Chain Neural Networks, *arXiv preprint arXiv:1805.00784* (2018)

[Blanchet 13] Blanchet, J., Gallego, G., and Goyal, V.: A markov chain approximation to choice modeling, *In 14th ACM Conference on Electronic Commerce* (2013)

[Blei 03] Blei, D., Ng, A., and Jordan, M.: Latent dirichlet allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022 (2003)

[Blei 06] Blei, D. and Lafferty, J.: Dynamic topic models, *Proceedings of the 23rd international conference on Machine learning*, pp. 113–120 (2006)

[Crisci 18] Crisci, A., Grasso, V., Nesi, P., Pantaleo, G., Paoli, I., and Zaza, I.: Predicting TV programme audience by using twitter based metrics, *In Multimedi a Tools and Applications*, Vol. 77, pp. 12203–12232 (2018)

[Figueiredo 16] Figueiredo, F., Ribeiro, B., Almeida, J., and Faloutsos, C.: TribeFlow: Mining & Predicting User Trajectories, *In: International Conference on World Wide Web*, pp. 695–706 (2016)

[Gambs 12] Gambs, S., Killijian, M., Cortez, D. P., and Miguel, N.: Next place prediction using mobility Markov chains, *Proceedings of the First Workshop on Measurement, Privacy, and Mobility*, pp. 1–6 (2012)

[HO 17] HO, C., Damien, P., Gu, B., and Konana, P.: The time-varying nature of social media sentiments in modeling stock returns, *Decision Support Systems* (2017)

[Iwata 12] Iwata, T., Yamada, T., Sakurai, Y., and Ueda, N.: Sequential Modeling of Topic Dynamics with Multiple Timescales, *ACM Trans. Knowledge Discovery from Data (TKDD)*, Vol. 5, No. 4, pp. 1–19 (2012)

[Khaing 17] Khaing, P. and New, N.: Adaptive methods for efficient burst and correlative burst detection, *In: 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)* (2017)

[Mitsuzawa 16] Mitsuzawa, K., Tauchi, M., Domoulin, M., Nakashima, M., and Mizumoto, T.: 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* (2016)

[Shen 07] Shen, Z.-J. and Su, X.: Customer Behavior Modeling in Revenue Management and Auctions: A Review and New Research Opportunities, *Production and Operations Management*, pp. 713–721 (2007)

[Spina 15] Spina, D., Trippas, J., Cavedon, L., and Sanderson, M.: SpeakerLDA: Discovering Topics in Transcribed Multi-Speaker Audio Contents, *Proceedings of the Third Edition Workshop on Speech, Language & Audio in Multimedia*, pp. 7–10 (2015)

[Wang 09] Wang, C., Blei, D., and Fei-Fei, L.: Simultaneous image classification and annotation, *Proc. CVPR* (2009)

[Wang 12] Wang, Y., Agichtein, E., and Benzi, M.: 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)

[小澤 13] 小澤 暁人, 吉田 好邦：マルコフ連鎖を用いた生活行動再現による家庭エネルギー需要の推定, 環境情報科学学術研究論文集, Vol. 27, pp. 97–102 (2013)

[本村 09] 本村 陽一：大規模データからの日常生活行動予測モデリング, *Synthesiology*, Vol. 2, No. 1, pp. 1–11 (2009)