Technical Report of the 1st Workshop on Latent Dynamics (Jun 16, 2010, Tokyo, Japan)

潜在トピックモデルを用いたデータマイニング

岩田具治*
Tomoharu Iwata

Abstract: 近年，文書や購買履歴などの離散データを解析する手法として， トピックモデルが注目されている． トピックモデルとは，文書が潜在意味（トピック）に基づいて生成される過程を確率的に表現したモデルである． トピックモデルを用いることにより，多様なデータに内在する隠れた構造を抽出できる．本稿では，基本となるモデルについて解説した後，トピックモデルの応用として，時間変化する購買履歴データの解析のためのトピック追跡モデルを紹介する．

Keywords: トピックモデル，生成モデル，ギブスサンプリング，購買行動解析

1 まえがき

近年，文書や購買履歴などの離散データを解析する手法として，bag-of-words 表現された文書の生成過程を確率的にモデル化したトピックモデルが注目されている．トピックモデルの代表例として，Probabilistic Latent Semantic Analysis (PLSA)[10] や Latent Dirichlet Allocation (LDA)[6] があり，情報検索 [10]，音声認識 [22]，可視化 [15]，画像認識 [25, 17]，推薦システム [11, 13] など，様々なデータマイニング分野に適用されている．トピックモデルの特徴は，一つの文書が複数のトピックの混合として表現されることである．一つの文書が一トピックで表される混合多項分布に比べ，トピックモデルは高い精度で文書をモデル化できることが確認されている [6].

2 トピックモデル

文書 d の出現単語集合を $\boldsymbol{w}_d = \{w_{dn}\}_{n=1}^{N_d}$ とする．ここで w_{dn} は文書 d の n 番目の単語，N_d は文書 d の単語数を表す．トピックモデルでは，各文書が固有のトピック比率 $\boldsymbol{\theta}_d$ を持ち，単語 w_{dn} は，$\boldsymbol{\theta}_d$ に従いトピック z_{dn} を選択した後，そのトピックに固有の単語分布 $\boldsymbol{\phi}_{z_{dn}}$ に従って生成される，と仮定する．文書集合を学習データとして推定したトピック比率 $\hat{\boldsymbol{\theta}}_d$ は，例えば，類似文書検索や文書分類 [6]，可視化 [12] に用いることができる．また，推定した単語分布 $\hat{\boldsymbol{\phi}}_k$ から，トピック毎に特徴的な単語を知ることができる．具体的には，トピックモデル (LDA) では，文書集合 $\boldsymbol{W} = \{\boldsymbol{w}_d\}_{d=1}^{D}$ は以下の過程で生成される．

(1) For each topic $k = 1, \cdots, K$:
  (a) Draw word distribution,
    $\boldsymbol{\phi}_k \sim \mathrm{Dir}(\beta)$,
(2) For each document $d = 1, \cdots, D$:
  (a) Draw topic proportion,
    $\boldsymbol{\theta}_d \sim \mathrm{Dir}(\alpha)$,
  (b) For each word $n = 1, \cdots, N_d$:
    (i) Draw topic,
      $z_{dn} \sim \mathrm{Mult}(\boldsymbol{\theta}_d)$,
    (ii) Draw word,
      $w_{nm} \sim \mathrm{Mult}(\boldsymbol{\phi}_{z_{dn}})$,

ここで K はトピック数，D は文書数，$\boldsymbol{\phi}_k$ はトピック k の単語分布，$\boldsymbol{\theta}_d$ は文書 d のトピック比率，z_{dn} は文書 d の n 番目の単語の潜在トピックを表す．また $\mathrm{Dir}(\cdot)$ はディリクレ分布，$\mathrm{Mult}(\cdot)$ は多項分布を表す．

トピックモデルにおける文書集合 $\boldsymbol{W}$ とトピック集合 $\boldsymbol{Z} = \{\{z_{dn}\}_{n=1}^{N_d}\}_{d=1}^{D}$ の完全尤度は下式で表される．

$$P(\boldsymbol{W}, \boldsymbol{Z}|\alpha, \beta) = P(\boldsymbol{Z}|\alpha)P(\boldsymbol{W}|\boldsymbol{Z}, \beta). \quad (1)$$

第一因子は $P(\boldsymbol{Z}|\alpha) = \prod_{d=1}^{D} \int P(\boldsymbol{z}_d|\boldsymbol{\theta}_d)P(\boldsymbol{\theta}_d|\alpha)d\boldsymbol{\theta}_d$ であり，$\{\boldsymbol{\theta}_d\}_{d=1}^{D}$ を積分消去することにより，以下の Polya 分布で表される．

$$P(\boldsymbol{Z}|\alpha) = \left(\frac{\Gamma(\alpha K)}{\Gamma(\alpha)^K}\right)^D \prod_d \frac{\prod_k \Gamma(N_{kd} + \alpha)}{\Gamma(N_d + \alpha K)}, \quad (2)$$

ここで $\Gamma(\cdot)$ はガンマ関数を表す．また第二因子も同様に Polya 分布，

$$P(\boldsymbol{W}|\boldsymbol{Z}, \beta) = \left(\frac{\Gamma(\beta V)}{\Gamma(\beta)^V}\right)^K \prod_k \frac{\prod_w \Gamma(N_{kw} + \beta)}{\Gamma(N_k + \beta V)}, \quad (3)$$

*NTT コミュニケーション科学基礎研究所，〒 611-0237 京都府相楽郡精華町光台 2-4，e-mail iwata@cslab.kecl.ntt.co.jp
NTT Communication Science Laboratories, 2-4, Hikaridai, Seikacho, Sorakugun, Kyoto

で表される．ここで V は語彙数である．

トピック集合 $\boldsymbol{Z}$ は，文書集合 $\boldsymbol{W}$ を入力とし，Collapsed ギブスサンプリング [9] を用いることで効率的に推定できる．文書 d の n 番目を生成する単語のトピック z_j，$j=(d,n)$，のサンプリング確率は下式により計算できる．

$$P(z_j=k|\boldsymbol{Z}_{\setminus j},\boldsymbol{W}) \propto \frac{N_{dk\setminus j}+\alpha}{N_{d\setminus j}+\alpha K}\cdot\frac{N_{kw_j\setminus j}+\beta}{N_{k\setminus j}+\beta V}, \quad (4)$$

ここで N_{dk} は文書 d におけるトピック k が割り当てられた単語数，N_{kw} はトピック k における単語 w の出現回数，$N_k=\sum_{k=1}^{K}N_{kw}$，$\setminus j$ は文書 d の n 番目の単語を除いたときの回数もしくは変数を表す．上式は，文書 d でのトピック k の割合と，トピック k での単語 w_j の割合の積で表されている．ディリクレ分布のパラメータ α および β は，不動点反復法 [19] を用いて完全尤度 (1) を最大化することによりデータから推定できる．例えば α は下式で更新される．

$$\alpha^{(\mathrm{new})} \leftarrow \hat{\alpha}\frac{\sum_d\sum_k[\Psi(N_{dk}+\alpha)-\Psi(\alpha)]}{K\sum_d[\Psi(N_d+\alpha K)-\Psi(\alpha K)]}, \quad (5)$$

ここで $\Psi(\cdot)$ はディガンマ関数 $\Psi(x)=\frac{\partial\log\Gamma(x)}{\partial x}$ を表す．全単語に対する潜在トピックのサンプリング (4) とパラメータの最適化 (5) を収束するまで繰り返すことによりモデルを学習できる．文書毎のトピック比率 $\boldsymbol{\theta}_d$ およびトピック毎の単語分布 $\boldsymbol{\phi}_k$ の推定値は下式により計算できる．

$$\hat{\boldsymbol{\theta}}_d=\frac{N_{dk}+\alpha}{N_d+\alpha K}, \quad (6)$$

$$\hat{\boldsymbol{\phi}}_k=\frac{N_{kw}+\beta}{N_k+\beta V}. \quad (7)$$

他の推論手法として変分ベイズ法 [6]，Collapsed 変分ベイズ法 [24]，期待伝搬法 (EP)[20]，パーティクルフィルタ [7] などが提案されている．文献 [2] では複数の推論手法の比較実験が行われている．

3 応用

トピックモデルは拡張性が高く，多様な情報を統合することを可能にする．例えば，著者 [21]，時間 [5, 26, 13, 14]，アノテーション情報 [3, 16] を統合したモデルが提案されている．

トピックモデルの一応用例として，時間発展する購買履歴データのためのトピック追跡モデル [13] を紹介する．トピック追跡モデルを用いることにより，ユーザの興味を予測し推薦システムやパーソナライズド広告に応用できるとともに，トピック毎の流行の時間発展を解析できる．購買履歴データにおけるユーザと商品は，文書データにおける文書と単語に対応する．つまり，時刻 t においてユーザ d が n 番目に購入する商品 w_{tdn} は，ユーザ固有のトピック比率 $\boldsymbol{\theta}_{t,d}$（興味を表す）に従ってトピック z_{tdn} を選択した後，トピック固有の商品分布 $\boldsymbol{\phi}_{t,z_{tdn}}$（流行を表す）に従って生成される．ここで，興味 $\boldsymbol{\theta}_{t,d}$ および流行 $\boldsymbol{\phi}_{t,k}$ は時間依存であることに注意．LDA ではこれらの多項分布パラメータは対称ディリクレ分布から生成されると仮定されているが，トピック追跡モデルではダイナミクスを考慮するために，過去のパラメータに依存するように拡張する．具体的には，興味の平均は，新たなデータが観測されない場合，その一時刻前の興味と同じであると仮定し，以下のディリクレ分布を興味 $\boldsymbol{\theta}_{t,d}$ の事前分布として用いる．

$$\boldsymbol{\theta}_{t,d}\sim\mathrm{Dir}(\alpha_{t,d}\hat{\boldsymbol{\theta}}_{t-1,d}), \quad (8)$$

ここで，平均は一時刻前の興味 $\hat{\boldsymbol{\theta}}_{t-1,d}$，精度（分散の逆数）は $\alpha_{t,d}$ である．精度 $\alpha_{t,d}$ は，直感的には，ユーザ d の時刻 $t-1$ と t 間での興味の一貫性を表す．興味の一貫性はユーザおよび時間に依存するため，精度 $\alpha_{t,d}$ を各ユーザ，各時刻でデータから推定する．精度を逐次推定することにより，変化する興味を柔軟に追跡できるようになる．興味と同様に，流行も一時刻前の興味に依存した以下のディリクレ分布から生成されると仮定する．

$$\boldsymbol{\phi}_{t,k}\sim\mathrm{Dir}(\beta_{t,k}\hat{\boldsymbol{\phi}}_{t-1,k}), \quad (9)$$

ここで平均は一時刻前の流行 $\hat{\boldsymbol{\phi}}_{t-1,k}$，精度は $\beta_{t,k}$ である．

トピック追跡モデルでは，新たに得られた購買履歴データと，過去に推定した興味・流行を用いて，現在の興味・流行を逐次的に推定する．すなわち，過去のデータはモデル推定に不要であり，保持する必要もないため，計算コストと記憶容量を低く抑えることができる．共役事前分布であるディリクレ分布を用いるため，ダイナミクスを考慮しない LDA と同様，Collapsed ギブスサンプリングによる効率的な潜在トピック推論が可能である．またハイパーパラメータである $\alpha_{t,d}$ や $\beta_{t,k}$ は，不動点反復法 [19] により完全尤度を最大化することによりデータから推定できる．

実購買履歴データを用いた実験により，トピック追跡モデルは，従来法に比べ購買行動をより高い精度で予測でき，かつ，大規模データでも効率的に扱うことができることを確認している．

トピック追跡モデルでは，LDA における事前分布に時間依存性を導入することで，時間変化する購買履歴データにも適用可能なように拡張している．文書データ，購買履歴データの以外にも，画像 [3, 8, 25]，ネットワーク [1]，音楽 [27] など，様々なデータでトピックモデル

の有効性が確認されている．その他のトピックモデルの発展として，ディリクレ過程を用いたトピック数の自動推定 [23]，トピック間相関の導入 [4]，トピック階層構造の導入 [18] などがある．

参考文献

[1] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.*, 9:1981–2014, 2008.

[2] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh. On smoothing and inference for topic models. In *UAI '09: Proceedings of the International Conference on Uncertainty in Artificial Intelligence*, 2009.

[3] D. M. Blei and M. I. Jordan. Modeling annotated data. In *SIGIR '03: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 127–134, 2003.

[4] D. M. Blei and J. Lafferty. A correlated topic model of science. *Annals of Applied Statistics*, 1(1):17–35, 2007.

[5] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *ICML '06*, pages 113–120, 2006.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[7] K. R. Canini, L. Shi, and T. L. Gri ths. Online inference of topics with latent Dirichlet allocation. In *AISTATS '09: Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, 2009.

[8] L. Cao and L. Fei-Fei. Spatially coherent latent topic model for concurrent object segmentation and classification. In *Proceedings of IEEE Intern. Conf. in Computer Vision (ICCV).*, 2007.

[9] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101 Suppl 1:5228–5235, 2004.

[10] T. Hofmann. Probabilistic latent semantic analysis. In *UAI '99: Proceedings of 15th Conference on Uncertainty in Artificial Intelligence*, pages 289–296, 1999.

[11] T. Hofmann. Collaborative filtering via Gaussian probabilistic latent semantic analysis. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 259–266. ACM Press, 2003.

[12] T. Iwata, K. Saito, N. Ueda, S. Stromsten, T. L. Griffiths, and J. B. Tenenbaum. Parametric embedding for class visualization. *Neural Computation*, 19(9):2536–2556, 2007.

[13] T. Iwata, S. Watanabe, T. Yamada, and N. Ueda. Topic tracking model for analyzing consumer purchase behavior. In *IJCAI '09: Proceedings of 21st International Joint Conference on Artificial Intelligence*, pages 1427–1432, 2009.

[14] T. Iwata, T. Yamada, Y. Sakurai, and N. Ueda. Online multiscale dynamic topic models. In *KDD '10*, 2010.

[15] T. Iwata, T. Yamada, and N. Ueda. Probabilistic latent semantic visualization: topic model for visualizing documents. In *KDD '08: Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 363–371. ACM, 2008.

[16] T. Iwata, T. Yamada, and N. Ueda. Modeling social annotation data with content relevance using a topic model. In *NIPS '09*, pages 835–843, 2009.

[17] L.-J. Li, R. Socher, and L. Fei-Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2036–2043, Los Alamitos, CA, USA, 2009. IEEE Computer Society.

[18] W. Li and A. McCallum. Pachinko allocation: Dag-structured mixture models of topic correlations. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 577–584, New York, NY, USA, 2006. ACM.

[19] T. Minka. Estimating a Dirichlet distribution. Technical report, M.I.T., 2000.

[20] T. Minka and J. Lafferty. Expectation-propagation for the generative aspect model. In *UAI '02: Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, pages 352–359, 2002.

[21] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *UAI '04: Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494, Arlington, Virginia, United States, 2004. AUAI Press.

[22] Y.-C. Tam and T. Schultz. Correlated latent semantic model for unsupervised language model adaptation. In *ICASSP '07: Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, volume IV, pages 41–44, 2007.

[23] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[24] Y. W. Teh, D. Newman, and M. Welling. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems*, volume 19, 2007.

[25] X. Wang and E. Grimson. Spatial latent Dirichlet allocation. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1577–1584, Cambridge, MA, 2008. MIT Press.

[26] X. Wang and A. McCallum. Topics over time: a non-Markov continuous-time model of topical trends. In *KDD '06*, pages 424–433, 2006.

[27] K. Yoshii, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno. An efficient hybrid music recommender system using an incrementally trainable probabilistic generative model. *IEEE Transactions on Audio, Speech and Language Processing*, 16(2):435–447, 2008.