4E1-4

ソーシャルネットワークにおけるリンク生成モデルとしてのLDAの提案

Proposal to use LDA as the Link Generation Model in Social Networks

倉持 俊也 Toshiya Kuramochi　土方 嘉徳 Yoshinori Hijikata　西田 正吾 Shogo Nishida

大阪大学大学院 基礎工学研究科
Graduate School of Engineering Science, Osaka University

Recently, many users have posted a large amount of text to Web using blogs or SNSs. Many researchers have continued to analyze that text because it represents users' latent interest. The LDA, a major topic model, is widely adopted as a method to extract users' interesting topics. Other hand, online social networks like Twitter is gaining interests by researchers. In Twitter networks, that is said follow relations represent users' interest. We propose the link generation model such as: (1) users have their own topic distributions, (2) follow relations are generated on the basis of their topic distributions. We can estimate these topic distributions by using the LDA. In this paper, we conduct LDA to the Twitter network and evaluate the results.

1. はじめに

近年，ブログや SNS が広く利用されるようになり，Web 上にはユーザが投稿した膨大な量のテキストが存在するようになった．これらのテキストはユーザの興味の対象を表した情報であるとして，様々な手法による解析が試みられてきた．中でも，トピックモデルの一種である LDA (latent Dirichlet allocation) [Blei 03] は，ユーザの潜在的なトピックを抽出する手法として広く用いられている．

一般に，LDA は文書集合中の単語に着目し，それぞれの文書の持つ潜在的なトピックを推定する手法である．LDA では，各文書に複数の潜在トピックが存在し，それらのトピックごとの単語の生成確率に基づいて単語が生成されるという文書生成モデルを想定している (図 1 (a))．LDA を用いてユーザのトピックを推定する場合，1 人のユーザが生成した全テキストを 1 文書と見なすことが多い．推定されたユーザのトピックの選択確率分布が，そのユーザの興味や嗜好を表している．

一方，ブログや SNS の分析では，各ユーザをネットワーク中のノードとしてモデル化することが可能であり，コミュニティ分析や情報拡散分析などに応用されている．近年，マイクロブログの一種である Twitter のネットワークを対象とした研究が広く行われ，Twitter のネットワークが社会的な知人関係を表したネットワークではなく，興味や関心の関係を表したネットワークであると報告されている [Kwak 10]．

我々もまた，オンラインソーシャルネットワークにおけるリンク (フォロー関係) は，そのユーザの興味に基づいて生成されると考える．つまり，それぞれのユーザは複数のトピックに対して興味を持っており，その興味に基づいてフォロー関係が生成されるというリンク生成モデルを想定する (図 1 (b))．図 1 に示すように，我々の提案するリンク生成モデルは，上述の文書生成モデルの文書をユーザに，単語を (フォロー対象の) ユーザに置き換えたものであるため，LDA によるトピック推定が可能であると考えられる．

本研究では，提案するリンク生成モデルに LDA を適用し，ユーザの興味のトピックを推定する．

連絡先: 倉持 俊也，大阪大学大学院 基礎工学研究科，大阪府 豊中市 待兼山町 1-3，06-6850-6383，kuramochi@nishilab.sys.es.osaka-u.ac.jp

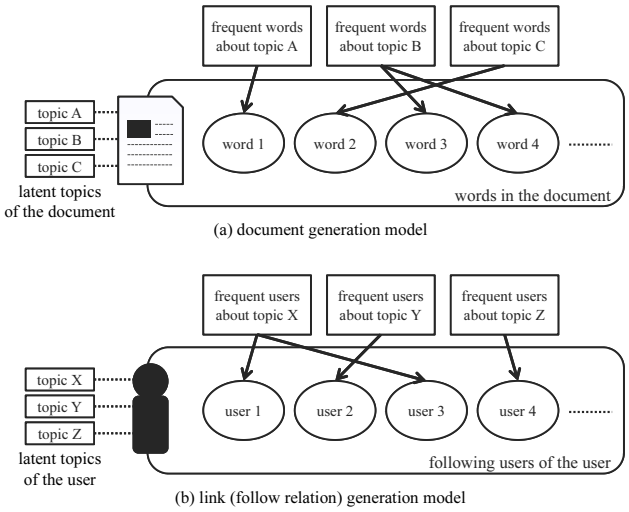

図 1: 文書生成モデルとリンク生成モデル

2. 関連研究

2.1 LDA の拡張と応用

Web 上に存在するユーザが投稿したテキストに対し，LDA を用いることで評判分析や自動要約，ユーザプロファイル作成などが行える．文書内で，あるトピックに関連する単語はまとまって出現することに着目し，文書内の窓レベルのトピックをモデルに組み込むことで，商品属性に着目したレビュー文の評判要約を可能にしたモデル [Titov 08] が考案されている．また，アイテムとユーザの特徴ベクトルを用いてレーティング値を予測するモデルに LDA が応用されている [Agarwal 10]．これらの研究は LDA の文書生成モデルを拡張することで，テキストから所望の情報を抽出している．我々は単語ではなくフォロー関係に着目したモデルを考案し，LDA を用いてユーザの興味のトピックを推定することを目的としている．

近年では，Twitter のテキストに LDA を適用した研究が増えてきている．影響力の強いユーザを発見するために，LDA を用いて推定したトピックを考慮した PageRank アルゴリズム [Weng 10] や，1 つのツイートに 1 つのトピックを割り当て

るように修正された LDA [Zhao 11] が提案されている．我々は Twitter ユーザの興味をユーザのフォロー関係に基づいて推定する．フォロー関係に対して LDA を適用した研究は，我々の知る限り存在しない．

2.2 ネットワーク生成モデル

複雑ネットワークの研究が進むにつれ，スモールワールド性やスケールフリー性のような特徴が明らかにされてきた．ネットワークの生成 (成長) 過程に着目した研究が多く為され，スケールフリー性を有するネットワークを生成するモデル (BA モデル [Barabasi 99]) が提案された．さらに，閾値モデル [Caldarelli 02] や最近隣接続モデル [Vázquez 03] など，スケールフリー性を説明するモデルが提案されている．これらのモデルは，ネットワーク全体を統一的な視点で捉え，ノードの次数やノード間の距離などに基づいてネットワークを生成するモデルである．本研究では，オンラインソーシャルネットワークにおけるエッジは，ネットワーク構造におけるノードの特徴に依存して生成されるのではなく，各ノードに該当するユーザの興味のトピックに基づいて生成されると考える．

3. 提案手法

LDA は文書集合中に出現する単語に着目し，それぞれの単語がどのトピックについて言及する単語であるか推定する手法である．本研究ではネットワーク上のリンクに対して，どのトピックに基づいて生成されているリンクであるか，同じ手法を用いて推定する．すなわち，我々は以下のようにユーザ u とフレンド $f_{u,n}$ の間にトピック $z_{u,n}$ のフォロー関係が生成されると考える．ただし，本稿ではあるユーザがフォローしているユーザをフレンドと呼ぶこととする．また，n はフレンドのインデックスである．

1. ユーザ u のトピック選択確率分布 θ_i に従い，トピック $z_{u,n} = k$ を生成
2. トピック $z_{u,n} = k$ のフレンド選択確率分布 ϕ_k に従い，フレンド $f_{u,n}$ を生成

全リンクのトピック $\boldsymbol{z}$ を推定する手法には，Gibbs サンプリングによる推定手法 [Grifiths 04] を用いる．この手法は，十分な反復回数が得られれば高い精度で推定が行えることが知られており，また実装も容易であるため広く用いられている．

4. 評価実験

Twitter Streaming API の sample エンドポイント*1を用いて 2013 年 3 月 21 日に取得したユーザのうち，言語設定を日本語としている 25,076 ユーザの全フレンドを取得し，トピック推定を行った．

図 2 に，ある 2 ユーザの推定されたトピック選択確率 (θ) を示す．ただし，トピック数は 200，推定の反復回数は 500 回である．トピック 41 の高頻度フレンド (フレンド選択確率 ϕ の大きいユーザ) は，主にあずまきよひこ氏，椎名高志氏，羽海野チカ氏，平野耕太氏ら漫画家であった．トピック 73 の高頻度フレンドは，まつもとゆきひろ氏，結城浩氏，徳丸浩氏，小飼弾氏らのようなプログラミングの解説書の著書らであった．なお，これらの 2 つのトピックを含む各ユーザが高い選択確率を示すトピックが，それぞれのユーザの興味を表していることはユーザ自身が確認している．

*1 http://dev.twitter.com/docs/api/1.1/get/statuses/sample

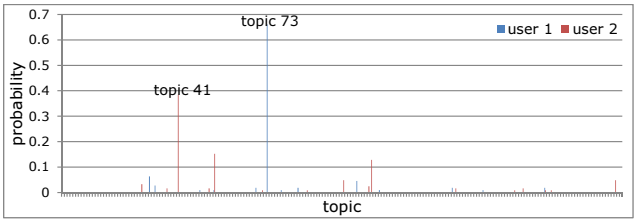

図 2: トピック選択確率

5. おわりに

本研究では，興味のトピックに基づいてフォロー関係が生成されると仮定し，LDA を用いてトピックを推定した．小規模な実験により，正しくトピック推定を行えることを確認した．今後は，ユーザのツイート (単語) から推定したトピックとの差異を被験者実験により明らかにする．また，友人推薦やユーザランキングなどのアプリケーション応用の可能性を調査する．

参考文献

[Agarwal 10] Agarwal, D. and Chen, B.-C., “fLDA: Matrix Factorization through Latent Dirichlet Allocation,” *Proc. of WSDM’10*, pp.91–100, 2010.

[Barabasi 99] Barabasi, A.-L. and Albert, R., “Emergence of Scaling in Random Networks,” *Science*, Vol.286, No.5439, pp.509–512, 1999.

[Blei 03] Blei, D.M., Ng, A.Y., and Jordan, M.I., “Latent Dirichlet Allocation,” *JMLR*, Vol.3, pp.993–1022, 2003.

[Caldarelli 02] Caldarelli, G., Capocci, A., Rios, P. De Los, Muñoz, M. A., “Scale-free networks from varying vertex intrinsic fitness,” *Phys. Rev. Lett.*, Vol.89, No.25, pp.258702, 2002.

[Grifiths 04] Grifiths, T. and Steyvers, M., “Finding scientific topics,” *PNAS*, Vol.101 (Suppl. 1), pp.5228–5235, 2004.

[Kwak 10] Kwak, H., Lee, C., Park, H., and Moon, S., “What is Twitter, a Social Network or a News Media?,” *Proc. of WWW’10*, pp.591–600, 2010.

[Titov 08] Titov, I. and McDonald, R., “Modeling Online Reviews with Multi-grain Topic Models,” *Proc. of WWW’08*, pp.111–120, 2008.

[Vázquez 03] Vázquez, A., “Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations,” *Phys. Rev. E*, Vol.67, No.5, pp.056104, 2003.

[Weng 10] Weng, J., Lim, E.-P., Jiang, J., and He, Q., “TwitterRank: Finding Topic-sensitive Influencial Twitterers,” *Proc. of WSDM’10*, pp.261–270, 2010.

[Zhao 11] Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., and Li, X., “Comparing Twitter and Traditional Media using Topic Models,” *Proc. of ECIR’11*, pp.338–349, 2011.