5P-07

Twitter 連携ニュースフィルタリングのための トピックモデルに基づくユーザの興味の学習

折本　伸之†　　渥美　雅保‡

創価大学大学院工学研究科情報システム工学専攻†　　創価大学理工学部情報システム工学科‡

1. はじめに

近年急速に普及し,注目を集めている代表的な SNS(Social Networking Service)として Twitter がある．Twitter では，ユーザはフォローと呼ばれる仕組みにより興味をもったユーザの最新のツイート(140 字以内の短文投稿)を常に受け取ることが可能である．ツイートには様々な情報が含まれており，これらの情報を抽出，活用することを目的とした研究は数多く行われている[1][2]．本研究では，ニュースサイトのフォローを対象として，それらのニュースサイトから投稿されるニュースツイートをユーザの興味からランク付けし，また，ツイートのリンク先のニュースを収集して LDA (Latent Dirichlet Allocation)[3]を用いたトピックモデリングによりユーザの興味を学習する方法に関して述べる．このユーザの興味の学習により，ユーザの興味に沿ったニュースツイートを優先的にユーザに提示する等，ニュースツイートフィルタリングの仕組みをアプリケーションに組み込むことが可能となる．

2. システムの構成

図 1 に本システムの構成を示す．本システムは，ツイートビューア，ウェブニュースクレイパー，コーパスと辞書ビルダー，ユーザの興味の学習器，及び興味モデルを用いたニュースツイートランキング器からなる．ツイートビューアはツイートの閲覧とそれらに対する興味の評価の機能を有する．ウェブニュースクレイパーは Twitter タイムライン上の Web ニュースツイートに対してニュースの文書をスクレイピングし，ニュースを収集する．コーパスと辞書ビルダーは，収集されたニュースデータをもとにコーパスと辞書を作成し，ユーザの興味学習器は LDA トピックモデルにより興味学習を行う．ニュースツイートランキング器はユーザの学習済み興味モデルに基づいてニュースツイートをランキングする．本論では，このうち，コーパスと辞書ビルダー，及びユーザの興味の LDA トピックモデル学習器について述べる．

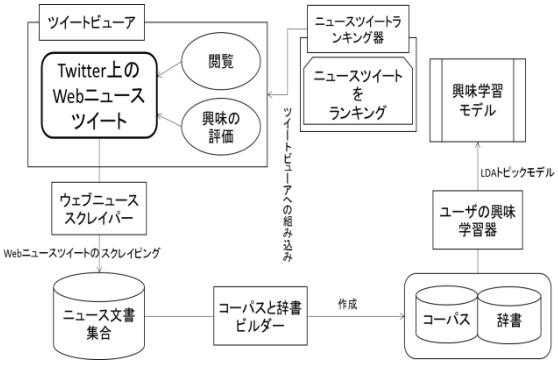

図 1. システム構成図

3. LDA トピックモデルによる興味の学習

LDA は，文書の確率的生成モデルで，文書をトピックの確率分布により，また，トピックを単語の確率分布により表すモデルである[3]．本研究では，LDA トピックモデルにより興味を表現する方法として，ユーザの興味を，ユーザが興味を示したニュース記事に高い確率で含まれるトピックにより表現する．

まず，Web ニュースツイートに対してニュース記事をスクレイピングし,記事を「文書」,「段落」,「文」の階層構造でデータベース化する．即ち，文書は複数の段落からなり，段落は複数の文から構成されるように管理する．次に，ニュース記事の文集合から辞書とコーパスを作成する．ここで，コーパスは，辞書に含まれる単語の Bag of Words(BoW)により表される．表 1 に辞書の例を，表 2 にコーパスの例を示す．そして，この辞書とコーパスに対して LDA を適用することにより，記事中の各文に含まれるトピックの分布，及び各トピックを特徴づける単語の分布を推定する．

段落のトピックは文のトピックから構成される．また，文章，即ちニュース記事のトピックは段落のトピックから構成される．段落と文書のトピック分布は，それぞれそれらに含まれる文と段落のトピック分布に対して，最大確率値が与えられた閾値を超える分布を足し合わせることにより構成される．ユ

表 1. 辞書の例

番号	単語	単語の出現回数
24	報告	41
36	このほど	16
75	大統領	57
619	通過	10
1272	知る	17

Users' Interest Learning based on Topic Models for Twitter Cooperative News Filtering
†Nobuyuki Orimoto
Graduate School of Engineering Dept. of Information Systems Eng., Soka University
‡Masayasu Atsumi
Dept. of Information Systems Sci., Faculty of Sci., and Eng., Soka University

表 2. コーパスの例

7:[(24:報告,1)(56:介入,1)(57:一蹴,1)(58:サイバー,1)(59:ニュース,1)(60:ロシア,1)(61:事実無根,1)]
115:[(210:男性,1)(244:続ける,1)(617:青信号,1)(619:通過,1)(624:ニューヨーク,1)(625:現れる,1)(626:走行,1)(627:中心,1)]
265:[(509:その後,1)(619:通過,1)(702:上院,1)(1218:可決,1)(1223:法案,1)(1225:賛成,1)(1227:多数,1)(1233:下院,1)]

ーザの興味がニュース記事に対して表示される場合，ユーザの興味はユーザが興味を示したニュース記事のトピック分布の集合により表現される．ユーザの興味を段落，もしくは文に対して特定できる場合は，ユーザの興味はユーザが興味を示した段落または文のトピック分布の集合により表現される．

4. 実験

4.1. 実験概要

Twitter タイムライン上に CNN.co.jp から投稿される 527 ツイートのニュース記事をスクレイピングし，記事を文章・段落・文に階層化したデータセットを作成した．527 記事に含まれる文の総数は 5306 文である．そして，データセットを基にコーパスと辞書を作成した．この辞書とコーパスに対して次の 2 つの実験を行った．

(1) トピックモデルの学習実験：この辞書とコーパスを用いて LDA トピックモデルの学習を行うことにより文集合のトピックモデルを生成した．この際，LDA 学習パラメータとして，トピック数を 100 から 1000 まで 100 ごとに変化させ，それぞれのトピック数でのパープレキシティを計算することにより，学習の成否とデータセットに対して適切なトピック数を評価した．

(2) ユーザの興味の評価実験：文のトピック分布から 3 で述べた方法によりユーザの興味のトピック分布を構成し，その分布がユーザの興味をとらえているかを定性的に評価した．

4.2. トピックモデルの学習実験の結果と考察

図 2 にトピック数を変えたときのパープレキシティの変化を示す．トピック数を増やすにつれてパープレキシティが減少していることから，トピック数の増加につれて学習がうまくいっていることが確認できた．また，これより，トピック数の設定としては 400 以上であればよいことも確認できた．

4.3. ユーザの興味の評価実験の結果と考察

図 3 に 3 で述べた方法により構成したユーザの興味のトピック分布の例を示す．図中，トピック名は，筆者が命名したものである．また，興味名もトピック分布から筆者が命名したものである．図 4 にこれら興味に高い確率で含まれるトピックの単語分布の例を示す．図 3，4 より，ユーザの興味の構造を推察できる．対象とする記事をさらに増やすことで，より多様なユーザの興味構造を獲得できると考える．

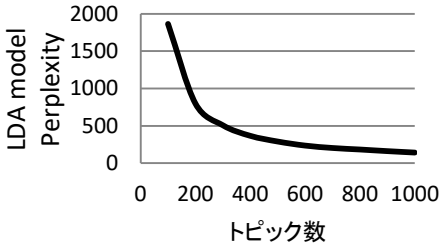

図 2. トピック数とパープレキシティ

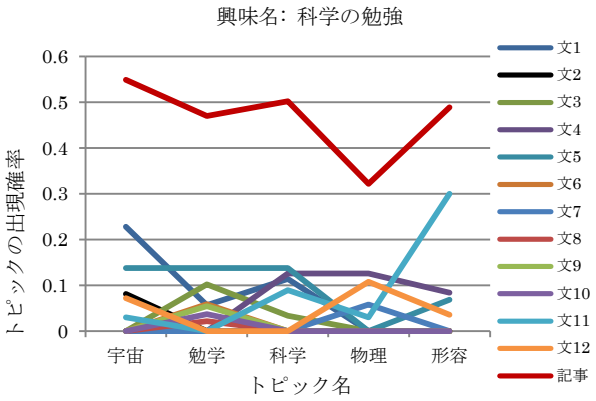

図 3. 興味のトピック分布表現の例

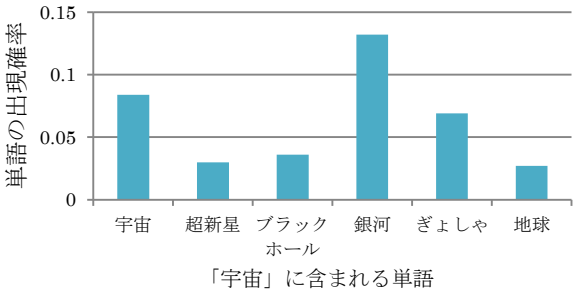

図 4. トピックの単語分布の例

5. むすび

本論では，LDA を用いたトピックモデリングによりユーザの興味を学習する方法について述べた．実験の結果，学習の精度向上と，ユーザの興味を反映した興味のトピック分布およびトピックの単語分布を得た．今後の課題として対象とするニュース記事を増やし，ユーザの興味モデルをより正確に評価し作成することがあげられる．

参考文献

[1] 近藤直人,内田理．Twitter を用いた LDA に基づくユーザの興味推定手法．言語処理学会第 21 回年次大会発表論文集, 2015.

[2] Keita Watanabe, Shohei Kato. Tweet Recommendation System Reflecting User Preference Based on Latent Dirichlet Allocation and Collaborative Filtering. The 28th Annual Conference of the Japanese Society for Artificial Intelligence, 2014

[3] David M. Blei, Andrew Y. Ng and Michael I. Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Research 3, pp.993-1022, 2003.