7S-01

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を用いたユーザの興味学習に基づくニュース Tweet ランキング

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 Tweet(140 字以内の短文投稿)を常に受け取ることが可能である．Tweet 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，ニュースサイトから投稿されるニュース Tweet をユーザの興味からランク付けするために，Tweet のリンク先のニュースを収集して LDA (潜在的ディリクレ配分法)[3]を用いたトピックモデルによって,ユーザの興味を学習する方法および，一連のニュースから興味のあるニュース Tweet をフィルタリングするためのランク付け手法に関して述べる．このユーザの興味の学習およびニュースフィルタリングのためのランク付けにより，ユーザの興味に沿ったニュース Tweet を優先的にユーザに提示する等，ニュース Tweet 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，Tweet ビューア，ウェブニューススクレイパー，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 Tweet ランキング器からなる．Tweet ビューアは Tweet の閲覧とそれらに対する興味の判定の機能を有する．ウェブニューススクレイパーは Twitter タイムライン上の Web ニュース Tweet 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 Tweet 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 Tweet 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，ユーザの興味の LDA トピックモデル学習器，及びニュース Tweet ランキング器について述べる．

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 Tweet 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．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．

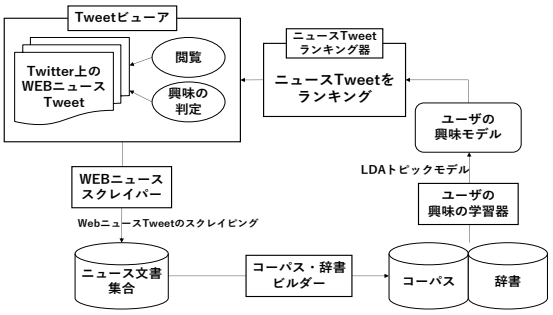

図 1　システム構成図

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. ニュース記事のランキング手法

3 で述べた方法により，ユーザが興味を持ったニュース文書集合から興味を表すトピックを学習して，ユーザの興味のトピック分布を生成する．そして，ユーザの興味のトピック分布と，ニュース記事のトピック分布の類似度を測定することで，ニュース記事をユーザの興味との近さによりランキングする. 分布間の類似度を測定するための距離尺度としては，L2 ノルム，KL ダイバージェンス，ヒストグラムインターセクションを用いる．

5. 実験

5.1. 実験概要

Twitter タイムライン上において，2018 年 9 月 15 日から 2018 年 12 月 26 日までの期間に CNN.co.jp から投稿された Tweet のうち，興味のある 1,628Tweet のニュース記事をスクレイピングし，記事を文章・段落・文に階層化した学習用データセットを作成した．1,628 記事に含まれる文の総数は 20,306 文である. そして学習用データセットを 407 記事ずつ 4 つに分割し，それぞれの学習用データセットを基にコーパスと辞書を 4 つ作成した．これらの辞書とコーパスを用いて次の 3 つの実験を行った．

① トピックモデルの学習実験：これら 4 つの辞書とコーパスを用いて, LDA トピックモデルの学習をそれぞれに対して行うことにより，記事のトピック分布を生成した．この際，LDA 学習パラメータとして，トピック数を 10

News Tweet Ranking based on User's Interest Learning using 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

から 100 まで 10 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 4 つの学習用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：記事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．

③ ニュース記事のランキング実験(1)：4 分割された学習用データセットに対し，興味のない記事をそれぞれ 407 記事ずつ追加し，新たにランキング用テストデータセットを 4 つ作成する．そして 4 分割された学習用データセットの各々に対して①②の方法で学習された 4 つのユーザの興味モデルを用いて，4 つのテストデータセットに含まれる記事のトピック分布を計算する．次に，これらの記事のトピック分布と，ユーザの興味モデルの間の類似度を 4 で述べた方法により計算し，テストデータセットの記事をランキングする．そして，4 つのユーザの興味モデルと 4 つのテストデータセットの 16 個の組み合わせの各々に対して，上位 20 記事に興味のある記事がどれだけ含まれているのかを評価する．

④ ニュース記事のランキング実験(2)：③の評価方法に加え，距離尺度を用いた類似度の測定結果に基づき，任意の距離の閾値より上位に興味のある記事がどれだけ含まれているのかを評価する．

5.2. 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，この結果より，興味モデルのトピック数の設定を 50 に設定した．

5.3. 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．

5.4. ニュース記事のランキング実験(1)の結果と考察

表 1 に，L2 ノルムを用いてランキングした際，上位 20 記事に興味のある記事がどれだけ含まれているかの割合を示し，表 2 に各距離尺度における各区間の興味ある記事の含有割合の平均を示す．表 1 の興味①②において，L2 ノルムは次区間の方が次々区間よりも興味のある記事をランキング出来ていることが確認できた．このことは表 2 より，KL ダイバージェンスについても同様のことが言える．しかし，同区間以外のランキング精度は，高くとも 0.65 であるため，興味のある記事が十分ランキング出来ているとは言えない．この原因の一つとして，分割サイズが大きくニュース記事の期間が長いことが考えられる．表 1，2 より，各距離尺度での同区間におけるランキング精度が高いことから，分割サイズをさらに小さくすることで，短期間におけるランキング精度が向上することが推察される．

5.5. ニュース記事のランキング実験(2)の結果と考察

表 3 にヒストグラムインターセクションを用いた際の各距離の閾値における次区間の興味ある記事の含有割合を示す．なお表中の括弧内の数値は，閾値以上に含まれている総記事数であり，括弧に隣接する数値は，その総記事数に含まれる興味ある記事の割合である．この結果から，興味①③においてはユーザの興味モデルを基に次区間の記事をランキングすることが出来たが，興味②では距離の閾値 0.55 以上において記事がランキングされなかった．これは興味②から③にかけての興味モデルが変化しており，ユーザの興味モデルを基にランキング出来なかったためと推察できる．5.4 と同様に，分割サイズを小さくすることで，より短期間のユーザの興味モデルを生成することで，ランキング精度が向上すると推察される．

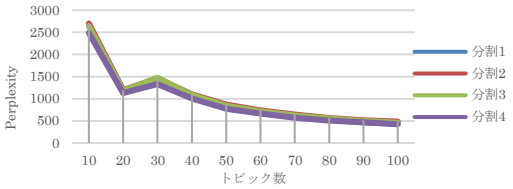

図 2 トピック数と Perplexity

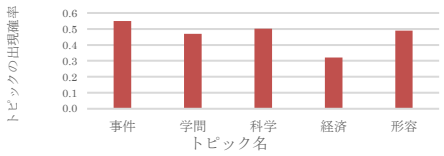

図 3 興味のトピック分布表現の例

表 1 L2 ノルムを用いた際の興味ある記事の含有割合

L2ノルム	興味①	興味②	興味③	興味④
分割①	1.00	0.40	0.35	0.70
分割②	0.60	1.00	0.45	0.60
分割③	0.55	0.65	0.95	0.45
分割④	0.60	0.55	0.55	0.95

表 2 各距離尺度における各区間の興味ある記事

	同区間	次区間	次々区間	将来全体
L2	0.975	0.600	0.550	0.583
KL	0.975	0.567	0.550	0.558
HI	0.963	0.517	0.600	0.550

表 1 ヒストグラムインターセクションを用いた際の各閾値における興味ある記事の含有割合

閾値	興味①	興味②	興味③
0.55	0.50(2)	該当なし	0.67(6)
0.50	0.61(23)	0.54(24)	0.50(26)
0.45	0.58(134)	0.54(127)	0.47(140)

6. むすび

本論では，LDA を用いたトピックモデルによりユーザの興味を学習する方法およびニュース記事のランキング手法について述べた．実験の結果，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．また，5.4 で述べた課題をはじめ，ニュース記事のランキング精度を向上させるための実験を継続する必要性がある．

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.