Twitter を用いた LDA に基づくユーザの興味推定手法

近藤 直人[1]，内田 理[2]

[1] 東海大学大学院工学研究科情報理工学専攻
3bdrm004@mail.tokai-u.jp
[2] 東海大学情報理工学情報科学科
o-uchida@tokai.ac.jp

1. はじめに

近年，ユーザの特徴を分析して有益な情報（パーソナライズされた情報）を提供するサービスが多数展開されている．例えば，Gunosy (http://gunosy.com/) は Twitter や Facebook への投稿内容を用いてユーザの特徴を推定し，その推定結果に基づいて情報を提供している．しかし，Gunosy が提供する情報はニュースサイトの記事やブログ記事などであり，動画や音楽といった情報は提供されていない．また，Yahoo! などの一般的な検索エンジンにおいては，Twitter などを用いて推定した個人の特徴に基づいた情報提供は行われていない．ユーザに有益な情報を提供するためには，SNS への投稿内容などから個人の特徴を推定し，それを利用してパーソナライズされた情報を提供することが望ましいと考える．しかし，SNS へのユーザの投稿内容を分析する際に単語の出現頻度にのみ注目すると，そのユーザが最も多く投稿した単語に偏って特徴が推定されてしまう．そこで本研究では，潜在的ディリクレ配分法（LDA）[1]と呼ばれる確率的トピックモデルを用いて，ツイートを対象とした分析を試みる．具体的には，対象ユーザと対象ユーザのフォロワーのツイートを用い，対象ユーザの興味を推定する手法を提案する．

2. 関連研究

古賀ら[2]は LDA を用いた Twitter 上のユーザ推薦システムを提案している．古賀らの研究では，各ユーザのツイートに対して LDA を適用し，類似ユーザの提示を行っているが，フォローをしたいと回答した被験者が 50%に達しておらず，改善の余地があるといえる．渡邊ら[3]は LDA と協調フィルタリングを用いて，ユーザの興味に基づいた情報推薦システムを提案した．LDA のみを用いた場合に比べて，精度が向上しているが，精度の面で改善の余地があると考えられる．

3. 提案手法

図 1 に提案手法の流れを示す．まず解析対象ユーザのツイートを取得する．次に，各ツイートに対して形態素解析を適用し，形態素に分割する．その後，LDA を適用することにより，各トピックにおける単語の確率分布に対して重みを付け，しきい値を超えたトピックをユーザの興味トピックとする．

3.1. Tweet の取得

解析対象ユーザのフォロワーは，対象ユーザ自身がツイートしていないが興味を持つ話題をツイートしていると考えられる．そこで，対象ユーザのみならず，対象ユーザのフォロワー（本研究では 49 人）のツイートも取得する（取得ツイート数は，各ユーザとも最新のもの 200 件とした）．本研究では，フォロワーの取得とツイートの取得に Twitter API v1.1 を利用した．フォロワーは，対象ユーザをフォローした日時が新しい順に 49 人のユーザを選択する．本研究では，1 人のユーザの全 200 ツイートを 1 文書とみなして処理する．以降の処理の前処理として，ツイート本文から改行やユーザ ID などを削除する．

3.2. 形態素解析の適用

文書の確率的なトピックモデルである LDA (Latent Dirichlet Allocation)を用いてツイートの解析を行うために，各文書に対して形態素解析を適用し，一般名詞と固有名詞を抽出する．本研究では，形態素解析器として MeCab を用いることとしたが，ツイート文には通常の形態素解析器では未知語と判断される単語が多いという特徴があるため，MeCab の辞書に Web 百科辞典の Wikipedia，はてはキーワード，及びニコニコ大百科のページタイトルを固有名詞として追加した．また，ニコニコ大百科のデータは，国立情報学研究所のダウンロードサービスにより，有限会社未来検索ブラジルから提供を受けたものを利用した[4]．各ページタイトルにはノイズとなりうるタイトルが含まれるため，以下に示す条件のタイトルを省いて辞書に追加した．

・一形態素
・一文字
・ひらがな，カタカナのみで 3 文字以下
・括弧を含む
・日付
・数字
・URL
・!?！？ のみ

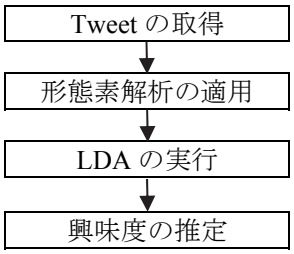

図 1　提案手法

3.3. LDA の実行

形態素解析を適用した文書群に対して LDA を実行する. LDA は 1 つの文書に対して複数のトピックが存在すると想定した確率的トピックモデルである. LDA によって, 各文書におけるトピック分布と各トピックを構成する単語の分布を算出することができる．本研究では，トピックの推定をより高い精度で推定できると報告されているギブスサンプリングを用いる[2][5][6]．トピック k おける単語 w の単語分布$\Phi_{k,w}$，文書 d におけるトピック k のトピック分布 $\theta_{d,k}$の算出式を以下に示す.

$$\Phi_{k,w} = \frac{N_{k,w} + \beta}{N_k + W\beta} \quad (1)$$

$$\theta_{d,k} = \frac{N_{d,k} + \alpha}{N_d + K\alpha} \quad (2)$$

ここで，αとβはハイパーパラメータ，Wは単語の種類数，Kはトピック数，N_kはトピック k に属する単語の数, $N_{k,w}$はトピック k に属する単語 w の数, $N_{d,k}$ は文書 d におけるトピック k に属する単語の数を表す.

3.4. 興味トピックの推定

LDA によって推定されるトピック分布は，対象ユーザのツイートに含まれる単語から推定されており，興味があっても投稿されていない内容は推定できない．本研究では，渡邊ら[3]が提案した協調フィルタリングを用いたものを拡張した手法と，tf-idf を用いた手法を提案し，性能の比較を行う.

3.4.1. 協調フィルタリング

協調フィルタリングは商品の推薦などに用いられる手法であり，オンラインショップの Amazon でも用いられている[7]．本研究では，ユーザベース協調フィルタリングを用いる．ユーザベース協調フィルタリングはあるアイテム群に対する評価が類似しているユーザ間で，あるアイテムに高い評価をしたユーザがいれば，他の類似しているユーザも高く評価する可能性が高いという仮定にも基づいている．本研究では, LDA のトピック分布を評価として協調フィルタリングを適用する．これにより，トピック分布が類似しているユーザのトピック分布によって，対象ユーザのトピックに関する評価を推定できる．まず, LDA によって算出されたユーザごとのトピック分布をもとに対象ユーザと各フォロワー間の類似度を計算する．ユーザ間類似度としてコサイン類似度を用いる．対象ユーザを u としたとき，あるフォロワーfとの類似度$cos(u,f)$を以下に示す.

$$cos(u,f) = \frac{\sum_{k=1}^{K}(\theta_{u,k} \cdot \theta_{f,k})}{\sqrt{\sum_{k=1}^{K} {\theta_{u,k}}^2}\sqrt{\sum_{k=1}^{K} {\theta_{f,k}}^2}} \quad (3)$$

式(3)に基づくトピック k における協調フィルタリングの適用式を以下に示す.

$$p_{u,k} = \overline{\theta_{u,k}} + \frac{\sum_{f=1}^{F}(\theta_{u,k} \cdot cos(u,f))}{\sum_{f=1}^{F}|cos(u,f)|} \quad (4)$$

ここで，$\overline{\theta_{u,k}}$は対象ユーザのトピック分布の平均，F はフォロワーの数である．式(4)より，協調フィルタリングを用いた対象ユーザ u のトピック k における推定評価値を以下に示す.

$$P_{u,k} = A\theta_{u,k} + (1-A)p_{u,k} \quad (5)$$

ここで$A(0 \le \mathrm{A} \le 1)$はパラメータであり，本研究では，協調フィルタリングの値に比重をおくために $A = 0.15$とした．また，生起確率が$1/K$より高いトピックを，対象ユーザが興味を持つトピックであるとした.

3.4.2. tf-idf

tf-idf とは文書における語の重要度を表す手法である. LDA によって算出されたトピック分布において，対象ユーザのツイートに含まれる単語に重みをおくことで，対象ユーザにとって興味のあるトピックが推定できると考えた．本研究では，対象ユーザのツイートに含まれる単語の重要度を算出するために tf-idf を用いる.

対象ユーザの各単語における tf-idf の値を算出する式を以下に示す.

$$tf\text{-}idf_{u,w} = \frac{N_{\mathrm{w}}}{N_u} \times \left(\log\frac{D}{D_w} + 1\right) \quad (6)$$

ここで, N_wは対象ユーザのツイートに含まれる単語 w の数，N_uは対象ユーザのツイートに含まれる単語数，D は全文書数，D_wは単語 w が含まれる文書数である.

tf-idf を用いた対象ユーザ u のトピック k における推定評価値$Q_{u,k}$を以下に示す.

$$q_{u,k} = \sum_{\mathrm{w}} \sqrt{\Phi_{k,w} \cdot tf\text{-}idf_{u,w}} \quad (7)$$

$q_{u,k}$においてすべてのトピックの和が 1 になるように正規化し，以下の式を適用する．

$$Q_{u,k} = B\theta_{u,k} + (1-B)q_{u,k} \quad (8)$$

ここで$B(0 \leq \mathrm{B} \leq 1)$はパラメータであり，本研究では tf-idf の値に比重をおくために$B = 0.3$とした．また，生起確率が$1/K$より高いトピックを，対象ユーザが興味を持つトピックであるとした．

4. 評価実験

4.1. 実験方法

提案手法の精度を検証するために，評価実験を 9 人に対して行った．それぞれのユーザに対して，対象ユーザと対象ユーザのフォロワー49 人のツイートを各 200 件取得する. 取得したツイートに対して，形態素解析を適用し, LDA を用いてトピック分布を求める. LDA のハイパーパラメータα, βは Zhao ら[8]に従い，$\alpha = \frac{50}{T}, \beta = 0.01$とした．また，トピック数は渡邊ら[3]に従い$K = 70$, ギブスサンプリング試行回数は 1000 回とした.

対象ユーザに LDA で生成された 70 トピックの単語集号を提示する．各単語集号は，生起確率上位 10 単語を提示した(生起確率が同一の場合は 10 単語を超えて提示した）．提案手法が対象ユーザにとって興味あるトピックを選択できるかを検証するために, 各トピックに対して“興味がある”，“興味がない”に分類してもらった（実験 1）．また，提案手法がそれぞれのトピックに対しての興味の強さを推定できているか検証するために，各トピックについての興味の強さを 4 段階（とても興味があるなら 4，全く興味がないなら 1)で評価してもらった(実験 2)．

4.2. 実験結果

実験 1 において，提案手法に加え，比較手法として LDA の生起確率を用いて興味推定した場合を含め，それぞれ適合率，再現率，F 値を求めた．協調フィルタリングを用いた結果を表 1，tf-idf を用いた結果を表 2，LDA による表 3 に示す．3 つの手法に対して，適合率，再現率，F 値の平均を比べたものを表 4 に示す.

協調フィルタリングを用いた場合と tf-idf を用いた場合の両方とも, LDA のトピック分布を用いただけの結果に比べ，適合率，再現率，F 値の全てにおいて高い精度を得た．提案手法はフォロワーのツイートより算出したトピック分布，単語分布を用いているため，対象ユーザ以外のツイートから興味トピックを推定する方法は有効であるといえる．協調フィルタリングと tf-idf の手法を比べると，協調フィルタリングの手法の方が高い精度を得られた．tf-idf を用いた手法は，対象ユーザのツイート中の単語に重点をおいているのに比べ，協調フィルタリングを用いた手法は，対象ユーザのフォロワーのトピック分布も用いている．対象ユーザのトピック分布のみでは興味があると推定されていなかったトピック，もしくは興味があると誤判定さてれてしまうトピックが減ったためであると考える．本研究では自動でツイートを行う bot に対する処理を加えていないため，ノイズとなる単語が様々なトピックに出現してしまった．また，被験者が興味ないと示したトピックに，辞書を拡張したために出現する意味のない単語が含まれていることがあった．そのために，対象ユーザが興味ありと回答したトピックと，提案手法によって興味ありと推定されたトピックとの間に差異が生じたと考える.

実験 2 において，それぞれの手法について平均 2 乗誤差 RMSE(Root Mean Square Error)を求めた. 平均 2 乗誤差を算出する式を以下に示す.

$$RMSE = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(x_k - \widehat{x_k})^2} \quad (9)$$

ここで, x_kはトピック k における被験者の評価, $\widehat{x_k}$はトピック k における各手法によって得られる興味の度合いであり，表 5 により求めた．実験 2 の結果を表 6 に示す.

表 6 より，いずれの手法においても興味の強さの誤判定は大きくないといえる．協調フィルタリングと tf-idf は，LDA に比べて誤差が小さくなった．この結果より，ユーザの興味があるトピックを推定するだけでなく，ユーザが生成された各トピックに対して興味をどれほどの強さで持つかを推定することも可能であると考える．また，実験 1 と同様に，ノイズとなりうる単語が様々なトピックに出現することが，対象ユーザの興味の強さと推定結果との間の差異を生じさせていると考える.

5. まとめと今後の課題

本研究では，対象ユーザと対象ユーザのフォロワーのツイートを用いた, LDA に基づく興味推定手法の提案を行なった．興味推定に協調フィルタリングや tf-tdf といった手法を用いることで，興味推定の精度が向上したといえる.

今後，精度を向上させる方法として，ストップワードやノイズとなりうる除外すべき単語の選定，自動でツイートを行う bot に対する処理を検討する．また，対象ユーザの興味トピックを選択するための重みを詳細に検討することで，精度の向上が図れると考える．興味推定されたトピック内の単語をクエリとして情報を提供する手法についても検討したい.

表 1　評価実験 1 の結果（協調フィルタリング）

	適合率	再現率	F 値
被験者 A	0.588	0.444	0.506
被験者 B	0.235	0.800	0.364
被験者 C	1.000	0.443	0.614
被験者 D	0.973	0.522	0.679
被験者 E	0.606	0.526	0.563
被験者 F	1.000	0.457	0.627
被験者 G	0.914	0.525	0.667
被験者 H	0.063	0.500	0.111
被験者 I	0.472	0.472	0.472
平均	0.650	0.521	0.512

表 2　評価実験 1 の結果（tf-idf）

	適合率	再現率	F 値
被験者 A	0.667	0.444	0.533
被験者 B	0.156	0.500	0.238
被験者 C	1.000	0.500	0.667
被験者 D	1.000	0.493	0.660
被験者 E	0.656	0.553	0.600
被験者 F	1.000	0.529	0.692
被験者 G	0.946	0.574	0.714
被験者 H	0.063	0.500	0.111
被験者 I	0.455	0.417	0.435
平均	0.660	0.501	0.517

表 3　評価実験 1 の結果（LDA）

	適合率	再現率	F 値
被験者 A	0.667	0.356	0.464
被験者 B	0.154	0.600	0.245
被験者 C	1.000	0.414	0.586
被験者 D	1.000	0.536	0.698
被験者 E	0.563	0.474	0.514
被験者 F	1.000	0.514	0.679
被験者 G	0.872	0.557	0.680
被験者 H	0.069	0.500	0.121
被験者 I	0.519	0.389	0.444
平均	0.649	0.482	0.492

表 4　評価実験 1 の結果（比較）

	協調フィルタリング	tf-idf	LDA
適合率	0.650	0.660	0.649
再現率	0.521	0.501	0.482
F 値	0.512	0.517	0.492

表 5　興味の強さの範囲

範囲	興味の強さ
$\theta_{u,k}, P_{u,k}, Q_{u,k} > 2/K$	4
$2/K \geq \theta_{u,k}, P_{u,k}, Q_{u,k} > 1/K$	3
$1/K \geq \theta_{u,k}, P_{u,k}, Q_{u,k} > 1/2K$	2
$1/2K \geq \theta_{u,k}, P_{u,k}, Q_{u,k}$	1

表 6　評価実験 2 の結果

	協調フィルタリング	tf-idf	LDA
被験者 A	1.165	1.062	1.128
被験者 B	1.429	1.429	1.502
被験者 C	0.727	0.707	0.918
被験者 D	1.254	1.189	1.231
被験者 E	0.765	0.737	0.828
被験者 F	1.352	1.293	1.493
被験者 G	1.363	1.298	1.612
被験者 H	1.378	1.378	1.326
被験者 I	1.134	1.128	1.121
平均	1.174	1.136	1.240

参考文献

[1] D. M. Blei, A. Y. Ng, M. I. Jordan, “Latent Dirichlet Allocation”, Journal of Machine Learning Research, Vol. 3, pp.993-1022, 2003.

[2] 古賀裕之, 谷口忠大, “潜在トピックに着目した Twitter 上のユーザ推薦システムの構築”, ヒューマンインタフェースシンポジウム 2010, 2010.

[3] 渡邊恵太, 加藤昇平, “トピックモデルと協調フィルタリングに基づくユーザ興味を反映した情報推薦システム”, 2014 年度人工知能学会全国大会, 2M3-4, 2014.

[4] 国立情報学研究所 情報学研究データレポジトリ, http://www.nii.ac.jp/cscenter/idr/index.html

[5] T. L. Griffiths, M. Steyvers, “Finding Scientific Topics”, Proceedings of the National Academy of Sciences, Vol. 101, pp. 5228-5235, 2004.

[6] 北島理沙, 小林一郎, “文書上の事象に基づいた潜在的トピック推定”, 2011 年度人工知能学会全国大会, 3H2-OS3-4, 2011.

[7] G. Linden, B. Smith, J. York, “Amazon.com Recommendations Item-to-Item Collaborative Filtering”, IEEE Internet Computing, Vol. 7, Issue 1, pp.76-80, 2003.

[8] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, “Comparing Twitter and Traditional Media using Topic Models”, Proc. 33rd European Conf. on Advances in Information Retrieval, pp. 338-349, 2011.