1A3-1

トピックモデルを利用したソーシャルテキスト上の同名他者推定

A Method for Classifying Homographs in Social Media Texts Using Topic Models

原田智彦　　津田和彦
Tomohiko HARADA　Kazuhiko TSUDA

筑波大学大学院システム情報工学研究科
Graduate School of Systems and Information Engineering, University of Tsukuba

The analysis of text data from social media is hampered by irrelevant noise data, such as homographs. The noise data are not usable, and the noise data make analysis, such as counting estimates, of the correct data difficult, which adversely affect the quality of the analysis results. We focus on this issue and propose a method to classify homographs that are contained in the text of the social media using topic models, and we report the results of the evaluation experiments.

1. はじめに

近年, ビッグデータを収集・分析して社会問題の解決, マーケティング戦略立案や業務改善などのビジネスに活かす取り組みが急速に拡がっており, ソーシャルメディアはその情報源のひとつとして注目されている. 代表的なソーシャルメディアである Twitter では, ユーザーは 140 文字以内のツイートと呼ばれるメッセージを使い, 日々の生活体験や思いを投稿できる. 投稿された情報は日常的に人から人へ伝わり, 多くのユーザーによってシェアされる. ツイートには, 購入した商品やサービスの選択基準や購入後の感想なども含まれるため, 企業にとって, ソーシャルメディア上の投稿情報から自社のビジネスに役立ちそうな投稿情報を収集・分析することの重要性が増している. 一方で,Twitter などソーシャルメディアのデータを対象とした研究や分析では, 検索結果や収集したデータ中に数多くのツイートが “ノイズ” として混在しているという共通の課題がある. これらのノイズは, 分析に役に立たないだけでなく, 分析結果の精度に影響を与える可能性がある. 例えば, 企業の評判分析の場合, 収集したツイート中に同じ名前の別の企業名が含まれていると分析精度が低下する要因となる. 本稿では, このツイート上での同じ名前を持つ企業や商品名の混在が引き起こす問題に着目し, トピックモデルを利用した同名他者推定の方法を提案し, 評価実験を行った結果について報告する.

2. トピックモデル

大規模かつ不均質な大量のテキスト情報から, 知識を獲得するための統計的モデリング方法の一つとして, 近年, トピックモデルが広く利用されている. トピックモデル [Hofmann 99] の特徴は,1 つの文書が複数のトピック情報の混合として表現されることである.1 つの文書が 1 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 [Blei 03]. 本稿では, トピックモデルとして良い性能を示すことが知られている潜在的ディリクレ配分法 (LDA; Latent Dirichlet Allocation)[Blei 03] を用いる.LDA では, 単語 w の集合を V とし, 単語 $w(\in V)$ の列によって表現された文書の集合とトピック数 K を入力として, 各トピック $z_k(k = 1, ..., K)$ における単語 w の確率分布 $P(w|z_k)(w \in V)$ および各文書 d におけるトピック z_k の確率分布 $P(z_k|d)(k = 1, ..., K)$ を推定する.LDA を用いることで, 文書中に現れる単語の出現確率を一様ではなく, 文脈に応じて変化すると考え, 単語 $w(\in V)$ の列によって表現された単語の集合とトピック数 K を入力として, 各トピック $z_k(k = 1, ..., K)$ における単語 w の確率分布 $P(w|z_k)(w \in V)$ および各文書 d におけるトピック z_k の確率分布 $P(z_k|d)(k = 1, ..., K)$ を推定することができる.

LDA を Twitter に対して適用した研究も多く報告されている [奥村 12]. Weng ら [Weng 10] は,LDA を用いて, 影響力のある Twitter ユーザーを検出する方法を提案している. Pennacchiotti ら [Pennacchiotti 11] は,LDA を元にしたツイート情報によるユーザーの分類モデルを提案している. 一方で, 通常, ツイートは手紙や報告書などに比べて短いため,LDA などの一般的なトピックモデルでは十分に意味を捉えることができないことが知られている. そのため,LDA を Twitter に対して適用する場合,1 ツイートを 1 文書とせず, 著者トピックモデル [Steyvers 04] の考えのもとユーザーの全ツイートを 1 文書として扱う方法が用いられる. これに対して,Zhao ら [Zhao 11] は,1 ツイートが 1 トピックであるという仮説を元に Twitter-LDA モデルを提案し, ツイートの短さによってトピックモデルが適切に推定できない問題を解消し, 前者のモデルと比べて優れていることを示している. また, 佐々木ら [佐々木 13] は, ユーザーの興味が日々変化することに対し,Twitter-LDA は従来の LDA と同様にツイートされる時間的な順序を考慮できない点に注目し,Twitter-LDA に Topic Tracking Model[岩田 10b] の機構を加え, ユーザーの興味と話題の時間発展を効率的にモデル化できる方法を提案している. 本稿では, 著者トピックモデルと同様に, ユーザーの全ツイートを 1 文書として扱い, ユーザーの興味分布をモデル化するが, ユーザーの興味と話題のダイナミクス (時間発展) の考慮を視野に入れ, 一定期間ごとのツイートを用いたトピックモデルの推定を行った.

3. 興味モデルを使った同名他者の推定

キーワード検索によって目的のツイート集合を収集しようとすると, 検索結果にはキーワードと同名の別の対象も含まれてしまう. 例えば,「アップル」というキーワードで検索を行い, コンピュータやディジタル家電メーカーの「アップル」に関するツイートを収集すると, 同じ「アップル」を冠した別の

連絡先: 原田智彦, 筑波大学大学院システム情報工学研究科, 東京都文京区大塚 3 丁目 29-1, s1230165@u.tsukuba.ac.jp

企業名や「アップルティー」や「アップルパイ」などフルーツの「アップル」も混入する. 通常, これらの目的と無関係なツイートはノイズとなるため, これらノイズを含んだ検索結果から, 目的のツイートを選り分けることが本稿の目的である.

本手法では, ツイート中の情報からツイートごとに目的のツイートかどうかを識別するのでなく, ツイートを投稿したユーザーがコンピュータやディジタル家電メーカーの「アップル」あるいはフルーツの「アップル」のどちらの発言をしやすいかをユーザーの興味分布によって識別する. そのため, 本手法は, まず, ユーザーが過去に投稿した複数のツイートの集合を 1 文書として扱い, LDA を用いて, ユーザーが各トピックに興味を持つ確率とトピックごとの単語の出現確率をモデル化する. このとき, 各ユーザーは固有のトピック比率 θ_u(ユーザー u が各トピックに興味を持つ確率を表す) を持つと仮定し, ユーザーがつぶやいた単語 w は θ_u に従いトピック k を選択した後, そのトピック k に固有の単語分布 ϕ_k に従って生成されたと考える. 次に,LDA のモデル推定によって得られた, 各ユーザーの興味分布 θ_u を機械学習の素性に用い, クラスタリングなどの手法により, ユーザーを分類する. 最後に, 収集したツイートを投稿したユーザーのクラスタにあわせてツイートを分類する.

4. 評価実験

実験は,2014/1/4~2014/1/11 に投稿されたキーワード「アップル」を含む日本語のツイートを使用し, 収集したツイートがコンピュータやディジタル家電メーカー「アップル」の製品やサービスに関するものか, フルーツなどそれ以外の「アップル」についてのものかを機械学習を使って識別する方法で行った.

実験データは,Twitter API を使って収集し, キーワードにマッチした 179,079 ツイートから 10,000 ツイートをランダムにサンプリングした. 次に,10,000 ツイートの中から,PR やボットを除外した上で, 各ツイートの投稿者について, 過去 1 年を遡ってツイートが収集できた 855 ユーザーによる 904 件を選択し, これにあらかじめ人手でコンピューターメーカーの「アップル」と「それ以外」の 2 種類の正解を付与し, テストデータを作成した.LDA の学習データは,855 ユーザーの過去 1 年間に投稿された 1,151,739 ツイートを収集し, ここから直近 1ヶ月分, 直近 3ヶ月分, 直近 6ヶ月分,1 年分の 4 セットの学習データを用意し, 語彙は一般名詞と固有名詞のみを抽出した. LDA の学習は,[岩田 10a] に倣い Collapsed ギブスサンプリング [Griffiths 04] を用い, また, ハイパーパラメータ α, β はサンプリングが行われるごとに不動点反復法により推定した. トピック数は $K = 150$[*1] を使用した.

次に,LDA モデルの学習で得られたユーザーごとの興味分布を素性に用い, 正解を付与したテストデータを使用して, 分類器による識別実験を行った. 分類器にはデータマイニングソフトウェア WEKA[*2] を利用し,10-fold cross validation によって評価した. 結果を表 1 に示す. なお, 表中の正解率は全正解に対する「アップル」と「それ以外」の正解数の割合で求めた.

実験の結果から, ユーザーごとの興味分布を素性に用いた識別は正解率で 75%を超える性能を示した. また,2 章で述べたように, ユーザーの興味と話題のダイナミクスを考慮すると, 比較的新しいデータのみを用いた方が識別性能が高くなると考えられるが, この影響は実験結果上で「直近 6ヶ月」と「1 年分」の違として見ることができる. 一方で,「1 年分」を除くと, 期間が長くなるほど反対に F 値が向上している. これは, 期間を長くすることで, ダイナミクスの影響よりも, 語彙数の増加が, トピックモデルを適切に推定する上でプラスに影響した可能性があると考えられる. ユーザーごとに十分な語彙数を集めることも課題である.

表 1: LDA による同名他者の識別実験

	直近 1ヶ月	直近 3ヶ月	直近 6ヶ月	1 年分
正解率 %	75.2	75.9	78.3	77.0
平均語彙数	651	1,769	3,512	6,698

*1 トピック数 K は事前実験による比較検討で perplexity 値によるモデルの安定性と処理時間の観点から決定した.

*2 アルゴリズムには事前実験による比較検討で高い性能を示した SMO を選択し, 他のオプションについてはデフォルトのままとした.

5. おわりに

本稿では, キーワード検索で収集したツイート集合にキーワードと同名の別の対象がノイズとして含まれてしまう問題に着目し, トピックモデルを利用して, 検索結果から目的のツイートを選り分ける方法を提案し, 評価実験の結果から提案方法の有効性を確認した. 今後は, ダイナミクスの取り入れたモデルによる効果の検証や精度向上に取り組む予定である.

参考文献

[Blei 03] Blei, D. M., Ng, A. Y., and Jordan, M. I.: Latent dirichlet allocation, *the Journal of machine Learning research*, Vol. 3, pp. 993–1022 (2003)

[Griffiths 04] Griffiths, T. L. and Steyvers, M.: Finding scientific topics, *Proceedings of the National academy of Sciences of the United States of America*, Vol. 101, No. Suppl 1, pp. 5228–5235 (2004)

[Hofmann 99] Hofmann, T.: Probabilistic latent semantic indexing, in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 50–57ACM (1999)

[Pennacchiotti 11] Pennacchiotti, M. and Popescu, A.-M.: A Machine Learning Approach to Twitter User Classification., in *ICWSM* (2011)

[Steyvers 04] Steyvers, M., Smyth, P., Rosen-Zvi, M., and Griffiths, T.: Probabilistic author-topic models for information discovery, in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 306–315ACM (2004)

[Weng 10] Weng, J., Lim, E.-P., Jiang, J., and He, Q.: Twitterrank: finding topic-sensitive influential twitterers, in *Proceedings of the third ACM international conference on Web search and data mining*, pp. 261–270ACM (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, in *Advances in Information Retrieval*, pp. 338–349, Springer (2011)

[奥村 12] 奥村 学：マイクロブログマイニングの現在, 電子情報通信学会第 3 回集合知シンポジウム (2012)

[岩田 10a] 岩田 具治：潜在トピックモデルを用いたデータマイニング, 第 1 回 Latent Dynamics Workshop, 2010/6 (2010)

[岩田 10b] 岩田 具治, 渡部 晋治, 山田 武士, 上田 修功：購買行動解析のためのトピック追跡モデル (人工知能, データマイニング), 電子情報通信学会論文誌. D, 情報・システム, Vol. 93, No. 6, pp. 978–987 (2010)

[佐々木 13] 佐々木 謙太朗, 吉川 大弘, 古橋 武：Twitter におけるユーザの興味と話題の時間発展を考慮したオンライン学習可能なトピックモデルの提案, 情報処理学会研究報告. MPS, 数理モデル化と問題解決研究報告, Vol. 2013, No. 3, pp. 1–6 (2013)