C-01

トピックモデルに基づく Twitter 上のアノーマルな日の検出

Detecting unnormality of tweets based on topic model

浜田 祥太郎† 角所 考† 岡留 剛†
Shotaro Hamada Koh Kakusho Takeshi Okadome

1. はじめに

人々が時と場所を選ばず web に接続し情報を容易に発信・収集することが可能になった昨今，急速に普及した web サービスとして twitter が挙げられる．その書き込み内容には社会での出来事による特徴がしばしば表れるため，twitter は新聞やニュースなどのマスメディアに替わる新しいメディアとしての側面がある．またマーケティングや災害防止など幅広い場面での活用の期待が高まっている．

そのため，twitter 上でとりあげられやすい話題の傾向や，実際の出来事が起こってから書き込みに反映されるまでの時間差，話題の移り変わり方などについて分析することは多くの意味を持つ．その前身となる課題のひとつとして，ある期間の twitter 上の書き込みが他の期間に比べて特徴的（アノーマル）であったことを知ることが挙げられる．

このことを踏まえ，本研究ではトピックモデルの考え方のひとつである Latent Dirichlet Allocation (LDA)[1]と確率分布間の距離を用いた文書のアノーマルスコア算出手法を提案し，それを用いて twitter における書き込み内容が特徴的な日（アノーマルな日）の検出を試みる．

本研究は，twitter という，明確なトピック性が存在せず，様々なトピックが混ざり合う文書においても，トピックモデルに基づいたアノーマリティ検出が可能であることを示している．

2．従来手法と提案手法

2.1 tf-idf

tf-idf は Salton らが 1983 年に提案し，情報検索などの分野で汎用に使われている手法である．単語ごとに tf-idf 値を持ち，ある文書におけるある単語の tf-idf 値が大きければ大きいほど，その文書中でのその単語の重要性が高いとみなされる．ここで，tf-idf 値は tf(Term Frequency)値と idf(Inverse Document Frequency)の積である(式 2.2)．ある単語の tf 値はその文書中におけるその単語の登場頻度である．ある単語の df(Document Frequency)値は文書集合の中でその単語が登場する文書の数であり，idf 値は df 値から式 2.1 のように計算される．直感的には，登場しにくい単語がある文書には多く登場している場合，その単語がその文書におけるキーワードである可能性が高いことになる．

$$idf = \log \frac{N}{df} \qquad N\text{は総文書数} \tag{2.1}$$

$$tf\text{-}idf = tf * idf \tag{2.2}$$

文書 d_i を，ベクトル $d_i=(tf\text{-}idf_{i1}, tf\text{-}idf_{i2}, \cdots, tf\text{-}idf_{in})$ で表し，ベクトル d_i と d_j のコサイン距離を計算することにより，それを文書 d_i と d_j の距離とする手法が広く用いられている．

$$D_{\cos}(p,q) = 1 - \frac{p \cdot q}{|p||q|} \tag{2.3}$$

2.2 提案手法

本稿で提案する，LDA と Hellinger 距離[2]を用いた文書間距離と，それらを用いたアノーマルスコアの定義について述べる．

2.2.1 LDA

LDA とは文書解析におけるトピックモデルの考え方のひとつで，トピックモデルとは文書のトピックやトピックに関連のある単語を文書集合から教師なし学習するための枠組みである．

LDA の考え方において，ある文書は複数のトピックの混合である．また，それぞれのトピックがすべての単語に対して出現確率を持つ．例えば，ある文書はスポーツと経済というトピックの混合である．そして，スポーツトピックにおいて「野球」は出現確率の高い単語であり，「虫」は出現確率の低い単語である．LDA において，文書内の単語ベクトル w は以下の生成プロセスに従うと仮定する．

1. Choose $N \sim \mathrm{Poisson}(\xi)$
2. Choose $\theta \sim \mathrm{Dir}(\alpha)$
3. For each of other of the N words w_n:
   (a) Choose a topic $z_n \sim \mathrm{Multinomial}(\theta)$.
   (b) Choose a word w_n from $p(w_n|z_n,\beta)$,
   a multinomial probability conditioned on the topic z_n.

また，図 2.1 は LDA のグラフィカルモデルである．

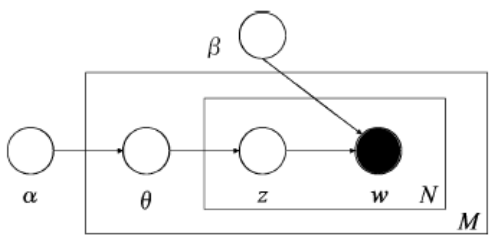

図 2.1 LDA のグラフィカルモデル．外側のプレートは文書ごとの試行を表し，内側のプレートは単語ごとの試行を表す．

†関西学院大学， Kwansei Gakuin University

LDA での学習により文書はトピック上の確率分布として表現される．このとき文書は単なる単語の集まり (bag-of-words) として扱われている．LDA で学習することで，文書を単語の種類数次元のデータから，潜在的トピック数次元のデータへと変換する．

2.2.2 文書間距離の計算

前節で，ひとつひとつの文書を潜在的トピックのディリクレ分布として表現した．この分布間の Hellinger 距離を文書間距離とする．Hellinger 距離は式 2.4 で表される確率分布間の距離尺度のひとつである．

$$D_H(p \| q) = \sum^{x} (p(x)^{\frac{1}{2}} - q(x)^{\frac{1}{2}})^2 \qquad (2.4)$$

2.2.3 アノーマルスコア

ある文書に関して，他のすべての文書との距離の合計をその文書のアノーマルスコアとする．

$$uns_d = \sum^{D} dist(d,i) \qquad D\text{は文書の数} \qquad (2.5)$$

ところで，従来の外れ値検知手法として知られているものとして，One-Class SVM[3]や LOF[4]がある．後述する実験においては，これらの手法との比較も行なう．

3．実験と考察

3.1 実験

提案手法を用いたアノーマルな日の検出と，比較のために tf-idf とコサイン距離を用いた手法について述べる．

3.1.1 提案手法の適用

Tweets の取得と bag-of-words 形式への変換

Twitter Streaming API を用いて tweets を取得した．その際，なるべく英語の tweets を集中して取得するため，以下の語を含む tweets のみを対象とした．

“the, my, me, our, you, he, him, she, her, they, it, this, that, is, are, have, has, had, do, dose, did, doing”

実験では，2013 年 11 月 25 日から 2014 年 4 月 15 日までの 142 日間（約 68 GB）の tweets を用いた．このとき 1 日分の tweets とは日本時間の 0 時から 23 時 59 分までの tweets を指す．

次に，tweets を文章の形式から，単語とその登場回数の対応表の形式に変換した．このとき，単語数が膨大になりすぎることを避けるために，1 日分の tweets に 10 回未満しか登場しなかった単語は無視した．また，”you”，“in” など文章の特徴に影響を与えないであろう stop word 約 2,000 語も無視した．こうして全 tweets に登場した単語の種類は約 22 万単語であった．

LDA での次元圧縮とアノーマルスコアの算出

前項で，約 22 万次元のデータとなった tweets を，LDA を用いて潜在的トピック数次元のディリクレ分布として表現する．ここでは 8，50，100 の 3 通りの潜在的トピック数について実験を行なった．これにより，各日の tweets が潜在的トピック数次元のディリクレ分布として表現される．

その後，前述の定義に沿って各日の tweets のアノーマルスコアを算出した．また，潜在的トピック数を 8 として次元圧縮した tweets については LOF 値の算出や，One-Class SVM でのアノーマリティ検出も行なった．なお，LOF 値の計算においても Hellinger 距離を用いている．

3.1.2 tf-idf とコサイン距離を用いた手法

比較のため，従来手法である tf-idf とコサイン距離を用いた尺度も計算した．

Tweets を tf-idf 値のベクトルとして表現し，提案手法において Hellinger 距離からアノーマルスコアを計算したように，ある日の tweets と別のすべての日の tweets とのコサイン距離の合計を計算し，アノーマルスコアの比較対象とした．

3.2 結果

ここまで，潜在的トピック数を 8，50，100 として LDA と Hellinger 距離を用いたアノーマルスコア，潜在的トピック数を 8 としての LOF 値，tf-idf ベクトルのコサイン距離を用いた尺度，単に日毎の tf-idf 値を用いた尺度について計算を行なった．以降，各尺度内のスコアの平均が 1 になるようにスケーリングした値を扱う．

なお，One-Class SVM を用いた外れ値検知も試みたが，異常ノードの割合を表す調整パラメータ ν をわずかに変えるだけで異常ノードの割合が大きく変わり，tweets の外れ値検知には適さないと判断した．

図 4.1 は潜在的トピック数を 8，50，100 として LDA と Hellinger 距離を用いて計算したアノーマルスコアのグラフである

図 4.2 は潜在的トピック数を 8 としたときのアノーマルスコアと LOF 値のグラフである．起伏が非常に似かよっている．

図 4.3 はコサイン距離を用いた尺度のグラフである．コサイン距離は各日の差異が少なく，大きなプロミネンスが見られない．

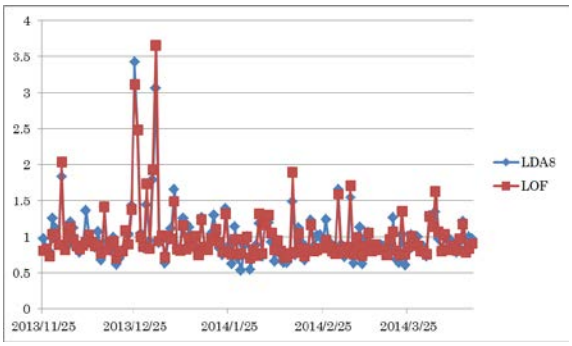

図 4.2 潜在的トピック数 8 の提案手法と LOF 値のグラフ．

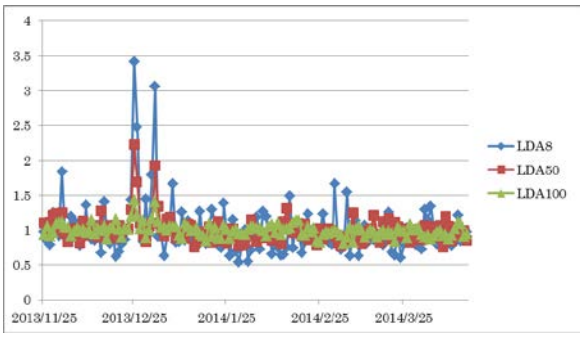

図 4.1 潜在的トピック数を 8, 50, 100 としたアノーマルスコア.

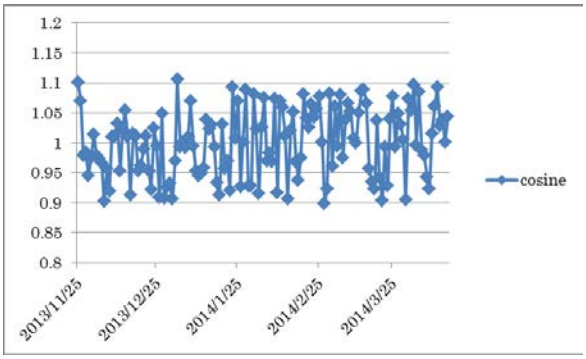

図 4.3 tf-idf とコサイン距離を用いた尺度のグラフ.

3.3 考察

図 4.1 から，提案手法の中では潜在的トピック数を 8 としたときが最もプロミネンスがはっきり表れていることが分かり，アノーマルな日の検出に適していると言える．以降，潜在的トピック数を 8 としたときの提案手法と他の手法を比較していく．

図 4.2 から，提案手法のグラフと LOF 値のグラフは大変似かよった起伏になっていることが分かる．このことから，tweets のアノーマルな日の検出においては，提案手法は LOF 値の近似として扱うことができることが分かる．

この実験で使用した tweets の期間である 2013 年 11 月 25 日から 2014 年 4 月 15 日の間で，twitter 上で多くの関連書き込みがされるイベントとして，クリスマスとニューイヤーズデイが挙げられる．これらのイベント時は tweets はアノーマルな内容となる．図 4.1 と図 4.2 において，提案手法や LOF を用いた尺度ではこれらのイベント時に大きなプロミネンスが見られ，これらのイベントを捉えられていることが分かるが，図 4.3 から分かるように，tf-idf とコサイン距離を用いた尺度ではこれらのイベントに起伏が見られず，全体的な起伏も小さい．

これらのことから，従来，文書の類似度計算などに用いられていた tf-idf とコサイン距離を用いた尺度では tweets のアノーマルな日をうまく検出できないが，一方で，LDA を用いて tweets を潜在的トピックの分布として表現し，それら分布間の距離を用いたアノーマルスコアや LOF 値を用いた手法であればアノーマルな日を検出できることが分かった．ここで興味深いことは，tweets は雑多な話題が混ざり合い人間にとって分かりやすいトピック性を持たないにも関わらず，トピックモデルに基づいたアノーマリティ検出ができるということである．

なお，提案手法において，クリスマスとニューイヤーズデイ以外でプロミネンスが見られた日として，2013 年 12 月 1 日，2014 年 1 月 7 日，2014 年 3 月 2 日などが挙げられる．これらのプロミネンスについて，tf-idf 値に基づくキーワード検出や実際の tweets 閲覧を元に，実世界の出来事との照らし合わせを行なったところ，以下のような実世界の出来事と関係していることが分かった．

- 2013 年 12 月 1 日
  俳優ポール・ウォーカーの事故死（キーワード:rippaulwalker）
- 2014 年 1 月 7 日
  カレッジフットボールの NCAA ディビジョンのチャンピオンを決定する BCS ナショナル・チャンピオンシップ・ゲームの開催（キーワード：bcscampionship）
- 2014 年 3 月 2 日
  子供向け専門チャンネルが主催する，アニメやスポーツ選手など全 16 部門の章を発表する子供たちの祭典，キッズ・チョイス・アワード(キーワード：kca)

4．まとめと今後の課題

本稿では，twitter の特徴を知り今後の活用についての議論を行なうために，トピックモデルに基づく文書間距離とそれを用いたアノーマルスコアを提案し，実際に twitter 上のアノーマルな日の検出を行なった．新聞記事などのように分かりやすいトピック性を持たない tweets においては，はっきりしたトピックの特徴を持たないものの，トピックモデルに基づきアノーマルな日を検出できることが分かった．また，これは従来の tf-idf とコサイン距離を用いた手法に比べ適格にアノーマリティを検出することができる．

次に，今後の課題について述べる．まず，本稿では潜在的トピック数を 8, 50, 100 としたが，妥当な潜在的トピック数の設定とその根拠については検証の余地がある．また，tweets を文書として扱う際に，時間の区切りや tweets された場所などに関して様々な条件を加える余地がある．本稿では英語の tweets を対象としたが，英語話者が様々な国に存在し，時差もあることなどから，世界的な規模の出来事でなければ影響が反映されにくい．そして，本研究で提案した文書間距離やアノーマルスコアの尺度，またアノーマルスコアに影響を与えた出来事の推定など，本来人の感覚により評価できる部分が多く，正確な妥当性の評価が難しいことも課題である．

参考文献

[1] Blei, D. : Probabilistic Topic Models, *COMMUNICATIONS OF THE ACM*, Vol.55, No.4, (2012), 77-84.

[2] C.M.ビショップ：**パターン認識と機械学習**,シュプリンガー・ジャパン,(2007).

[3] Scholkopf, B., J. Platt, J. Taylor, A. Smola, and R. Williamson : Estimating the Support of a High-Dimensional Distribution, *Neural Computation*, Vol.12, No.5, (2000), 1207-1245.

[4] Breunig, M., H. Kriegel, R. Ng, J. Sander : LOF: Identifying Density-Based Local Outliers, *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, (2000), 93-104.