

Web ニュース記事によるツイート分類に関する一検討

今井克真 †

小林亜樹 ‡

† 工学院大学工学部情報通信工学科

1 はじめに

文書分類技術の一つとして、文書データによる LDA トピックモデルが用いられている。しかし、文書長が短く、語彙にも偏りがあるとされるツイートに対しては、あまり有効ではないとされてきた。山田ら [1] のように Twitter 上で告知されたイベントを基にイベント名称、開催期間、開催場所といったイベントの情報を抽出するものがあるが、これは Twitter 上で告知されていないイベントに関しては抽出できない。そこで本論文では、Web ニュース記事で言及されている物事（ここではカテゴリと呼ぶ）に対して、ツイートがどのカテゴリへの言及であるかを判別するという観点において、Web 記事本文を 1 文書とし、名詞で構築した LDA トピックモデルに基づき、ツイート単独でのトピック分布と記事単独でのトピック分布の類似度からツイートのカテゴリ分類した。

2 提案手法

2.1 概要

カテゴリ分類した Web ニュース記事を抽出する。抽出した 1 記事本文を 1 文書として、記事本文を MeCab で形態素解析した結果得られた名詞のみの Bag of Words を学習させ、LDA トピックモデルを構築する。トピックとは単語その単語の生起確率の組の集合のことである。構築した LDA トピックモデルに対して、Web ニュース記事とツイート本文の学習した名詞 Bag of Words を用いたツイートから、記事-ツイートのトピック分布のコサイン類似度を計算する。

トピックを構成する語は Web ニュース記事中に出現した名詞のみとしているため、ツイートのみに出現する名詞は考慮されない。

2.2 LDA トピックモデルの構築

本論文では文書分類技術として、トピックモデルである LDA を使用する。LDA の実装には Python の gensim を用いた。学習データ文書を形態素解析し得られた名詞のみ LDA に学習させる。トピック数 K を指定し、LDA を適用して LDA トピックモデルを構築する。得られた LDA トピックモデルに対して、学習に使用した各 Web ニュース記事本文と分類を行いたい各ツイート本文そ

れぞれを構成するトピックの確率分布であるトピック分布を求め、その類似度を計算する。類似度の計算にはコサイン類似度を用いた。ベクトル \mathbf{D} , \mathbf{T} のコサイン類似度は (1) 式で計算される。

$$\cos(\mathbf{D}, \mathbf{T}) = \frac{\mathbf{D} \cdot \mathbf{T}}{\|\mathbf{D}\| \|\mathbf{T}\|} \quad (1)$$

類似度最大となった Web ニュース記事のカテゴリとして分類を行う。類似度最大となる記事が複数あった場合は、カテゴリ毎の類似度最大の記事数 $C_{x|x=a, \dots, e}$ が最も多いカテゴリに、カテゴリ別で C_x が同数の場合は複数のカテゴリに属すると判断する。

3 実験

3.1 目的

Web ニュース記事とツイートとの類似度により、選挙に関するツイートのうちどれだけのツイートが選挙カテゴリに分類されるかの実験をいくつかのトピック数で行い、トピック数によっては選挙に関係するツイートの多くが選挙カテゴリに分類され、選挙に関係しないツイートの多くが選挙カテゴリに分類されないか調べる。

3.2 Web ニュース記事

Web ニュース記事として、読売新聞の web ページ YOMIURI ONLINE* のニュース記事からテキストを収集した。収集したテキストは衆議院選挙に関する 20 記事 $\{D_{a1}, \dots, D_{a20}\}$ の本文、テレビに関する 20 記事 $\{D_{b1}, \dots, D_{b20}\}$ の本文、同様に竜王戦、就活、東京オリンピック・パラリンピックに関する記事をそれぞれ 20 記事ずつ $\{D_{c1}, \dots, D_{c20}\}$, $\{D_{d1}, \dots, D_{d20}\}$, $\{D_{e1}, \dots, D_{e20}\}$ の本文の合計 100 テキストである。

web から html を取得するのに Python の urllib を、html からテキストを取得するのに html2text を使用した。

3.3 ツイート

ツイートは 4 カテゴリ計 400 件用意した。うち 2 つは「衆議院」を含むもので、2017 年 10 月 10 日のツイートである。人手で選挙に対する感想意見を含むと判断したツイート 100 件 $\{T_{a1}, \dots, T_{a100}\}$ と、含まないと判断したツイート 100 件 $\{T_{b1}, \dots, T_{b100}\}$ の 2 カテゴリである。3 つ目は、同日のツイートから乱択した 100 件

Fundamental Analysis on Categorizing Tweets into Relevant Web News Article.

†Katsuma Imai †Aki Kobayashi

‡Department of Information and Communications Engineering, Faculty of Engineering, Kogakuin University

*<http://www.yomiuri.co.jp/election/shugin/?from=yenav1>

$\{T_{c1}, \dots, T_{c100}\}$ で選挙に関係すると判断されるツイートはなかった。4つ目は、別期間の「江東花火大会」を含む100件 $\{T_{d1}, \dots, T_{d100}\}$ で、Web ニュース記事カテゴリにはないカテゴリとして用意した。

3.4 LDA トピックモデルの構築

形態素解析には MeCab 0.996, 辞書は ipadic を使用した。Web ニュース記事 $\{D_{xy}|x \in \{a, \dots, e\}, y \in \{1, \dots, 20\}\}$ 1 記事を 1 文書とし、形態素解析し得られた名詞のみ 4292 単語からコーパスを作成し、LDA に学習させる。トピック数 $K = \{5, 10, 15, 20, 30, 50\}$ として LDA トピックモデルを構築する。

3.5 ツイート分類

構築した LDA トピックモデルに対して、各 Web ニュース記事 $\{D_{xy}|x \in \{a, \dots, e\}, y \in \{1, \dots, 20\}\}$ と各ツイート $\{T_{xy}|x \in \{a, \dots, d\}, y \in \{1, \dots, 100\}\}$ それぞれのトピック分布を求め、コサイン類似度を計算する。

1 ツイート中で使用された名詞数の平均は 25, そのうち学習に使用された名詞数の平均は 14 であった。ツイートとのカテゴリ推定は類似度が最大となった Web ニュース記事のカテゴリとして分類を行う。

3.6 実験結果

$K=30$ における選挙に対する感想意見を含むと判断したツイートを 100 件に対する結果を図 1 に示す。縦軸の上から順にテレビカテゴリ記事 D_{a1} から D_{a20} , 竜王戦カテゴリ記事 D_{b1} から D_{b20} , 就活カテゴリ記事 D_{c1} から D_{c20} , 東京五輪カテゴリ記事 D_{d1} から D_{d20} , 選挙カテゴリ記事 D_{e1} から D_{e20} を並べてある。横軸には選挙に対する感想意見を含むと判断したツイート T_{a1} から T_{a100} をカテゴリ分類の結果でソートした順に並べてある。

1 ツイートと 1 記事すべての組の類似度を計算し、ツイートに対し記事の類似度が最大値をとればマスに色を付け、分類結果のカテゴリ毎にも分類された範囲に色を付けた。図 1 からツイート T_a の多くが選挙カテゴリと判断できた。100 ツイート中 68 ツイートが選挙カテゴリ、31 ツイートが東京五輪カテゴリ、5 ツイートが就活カテゴリ、1 ツイートが竜王戦カテゴリと分類された。このうち 5 ツイートは選挙カテゴリと東京五輪カテゴリの 2 カテゴリに属すると分類された。

「衆議院」を含むが感想意見を含まないとしたツイート集合では、59 件が選挙カテゴリに分類された。乱択ツイート集合では、38 件がテレビカテゴリ、47 件が就活カテゴリであった。花火大会ツイート集合は、80 件が就活カテゴリに分類された。

3.7 考察

選挙の感想を含むとしたツイートの分類では、複数のカテゴリに属したツイートは 5 ツイートとも D_{d20}

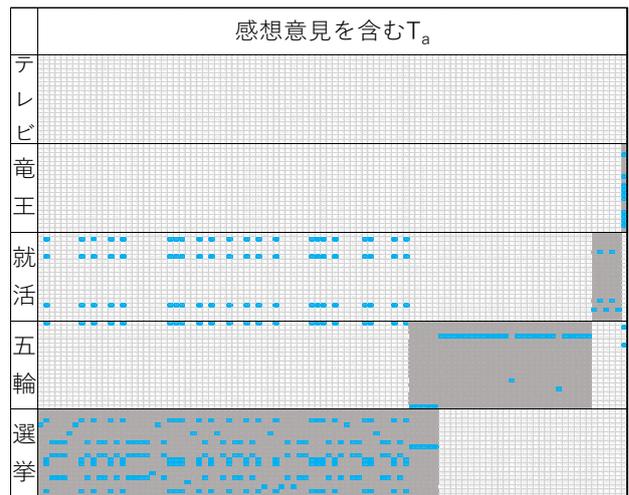


図 1: ツイート T_{a1}, \dots, T_{a100} との類似度

と D_{e9} の 2 つの記事との類似度が最大となった。記事 D_{d20} と D_{e9} はいずれもトピック 3 のみのため類似度が 1 であったためである。選挙の感想を含まないとしたツイートもほぼ同精度で選挙カテゴリに分類されており、このような意味内容の違いは分類に影響しないと考えられる。

乱択、花火ツイートも、特定のカテゴリに分類される傾向にあったが、これは単純に類似度最大記事のカテゴリへと分類しているためである。正解カテゴリがないツイートの各記事との類似度は全般に低く、何らかの閾値処理を行うことで適切な分類結果（いずれのカテゴリにも分類されない）を得ることができると考えられる。具体的な閾値の設定などは今後の課題である。

4 結論

本論文では、Web ニュース記事とツイートとの LDA トピックモデル上での類似度による、ツイートのカテゴリ分類手法を提案し、実験により一定の精度が得られることを確認した。分類基準に類似度の閾値による判定を加えるなどした上で、分類性能の検証を進めていく予定である。

謝辞

本研究の一部は科研費 (26242013) の助成を受けたものである。

参考文献

- [1] 山田 渉, 菊池 悠, 落合 桂一, 鳥居 大祐, 稲村 浩, 太田 賢, “マイクロブログを用いたイベント情報抽出技術”, 情報処理学会論文誌, Vo57, No. 1, pp.123-132, Jan. 2016.