トピックモデル用いたSNSにおける少数意見文の推定

Estimating Minority Opinion Sentences in Social Networking Service Using Topic Model

譚修[1)]
指導教員　亀田弘之[2)]

1) 東京工科大学大学院　バイオ・情報メディア研究科　思考と言語研究室
2) 東京工科大学　コンピュータサイエンス学部

Keywords：Minority Opinion　SNS　Topic Model

1．はじめに

SNS 時代になり、社会事件に関わるディスカッションは誰にも参加できる。短時間では数万件投稿を超える大規模なディスカッションも顕している。この状況に至っては、大量意見が手に入る。また、サイバーカスケートにより、SNS 利用者は同じ考えを持ち同士を結びつけることを簡易にする特徴もある。大量意見の中では、類似感想・論点が多くをしめ、同調現象が起きる。「みんな」以外の立場による意見を排除され、遮断する。

本研究では、社会事件に関する客観的世論の把握を目指す。これため、各々小さな意見が見つけることは解決するべき課題である。

2．関連研究

大量テキストデータから少数者の意見を抽出及び分析する研究は多くある。単語の情報量より少数意見に含める重要文（ハイライト）の抽出［1］や、TF-IDF と N-gram を用いたツィートデータからツイートの傾向及び少数意見の抽出[2]がある。また、SNS 上でのユーザ行動を分析し、少数派の分布研究［3］したものがある。しかし、抽出した少数意見に雑多な話題を含み、論点の判断が困難な意見文も多く見られる。そこで本研究では、Twitter のツイートを対象とし、ユーザが投稿した社会事件に関す意見を、LDA（Latent Dirichlet Allocation）によって機械的に再現できるかを検証する。その上で投稿数が少ないトピックから少数意見の推定を試みる。

3．提案手法

LDA とは自然言語処理における文書分類手法の一種である。文書における単語の出現頻度を推定し、似たような単語出てくる類似トピックを把握できるというモデルである。

そこで、Twitter 上の投稿した意見文は複数トピックを持つと仮定する。複数トピックを抽出するために、以下の手法((I)~(IV))を提案する。

(I). 意見文データを用意
(II). 形態素解析よって意見文を単語に分け
(III).意見文のベクトル化（テキストの TF-IDF 値）
(IV).LDA に導入

4．実験

4.1. 意見文データの抽出

分析事例は、ソフトバンクグループの孫正義社長が新型コロナウイルス対策として、「PCR 検査」を 100 万人分無償で提供すると発言し、Twitter 上で非難が寄せられた。事件の起こった日を含む 2020 年 3 月 11 日から 2020 年 5 月 15 日まで孫正義社長に関連するツイートを収集した。検索クエリ”孫正義 OR PCR 検査”に該当するツイート 5115 件を対象とした。Web クローラを利用しユーザ名、URL、ツイートテキストを取得した。

4.2. 意見文データ前処理

ベクトル化の対象は、名詞、動詞、形容詞とした。URL や特殊記号を除去する。日本語が解析場合にはテキストを単語ごとに区切り、その間に空白を置く必要がある。今回は、URL と特殊記号に対する解析精度が高いとされた形態素解析ツール Nagisa を用いる。

4.3. 意見文のベクトル化

処理したツイートテキストを対象とし、TF-IDF を用いた単語に重み付きをする。また、精度を上げるため、自分で作成した 1000 語ほどストップワードの除去も行う。

4.4. トピックの構築

ベクトルを基に LDA モデルを構築する。scikit-learn の LDA パッケージを用いる。

5．結果と評価

まず、トピックの分布状況を図 1 に表す。

LDA によってトピックを再現した一部の結果は図 2 に示す。図には Topic1 に関するツィート数が全体の約 22.7%を占める。Topic18 は 1.4%、Topic 19 は 1.3%、Topic 20 は 1.3%、Topic 21 は 1.1%を占める。ここで、Topic 18～21 はこの事件に対する少数意見のトピックとした。Topic18 に関するツイートには、「本当にそう思います。 PCR 検査 100 万人やってほしかったです…医師会と組んで都市部の抗体検査 100 万人はできませんか？抗体検査は PCR 検査と比べて技師の負担やかかる時間も少ないようです。」がある。

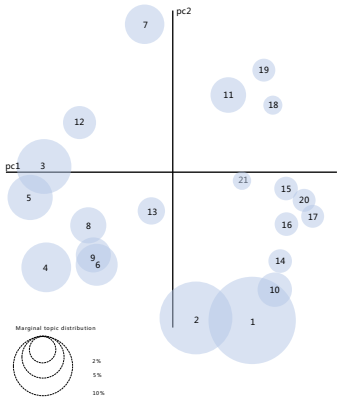

図 1　Topic 1 ～21 の分布

Topic1	日本,感染,キット,韓国,思う,言う,コロナ	22.7%
…	…	
Topic18	抗体,キット,医師,出る,コロナウイルス,病院,欲しい,現場,新型,医療,責任,出来る,正しく,思い,用意,全国	1.4%
Topic19	医療,崩壊,100 万,イタリア,提供,ハンク,ネット,ソフト,やめろ,無料,反応,マスク,撤回,困難,否定,入手	1.3%
Topic20	コロナウィルス,sb,対応,早期,大事,ウィルス,多数, covid,死者,軽症,変え,破壊,基準,増加,インフル,大切, 分析,発見	1.3%
Topic21	無償,提供,ニュース,100,撤回,医療,ライン,時間,新聞, 崩壊,読売,批判,経済,新型,ツイート,コロナ,簡易,リフライ	1.1%

図 2　孫正義ツイート炎上事件トピック単語分布

6．まとめ

本研究では、Twitter 上発生事件を対象として事件に関するツイートを抽出し、また、LDA を用いたトピックを再現し、全体に占める割合が少ないトピックから少数意見を抜き出して見ていく。ただ、あるトピックが他のトピックと重複する状況も発生した。今後は、トピック数の設定を慎重に検討する必要がある。

参考文献

[1] 松尾 豊，篠田 孝祐，石塚 満: 電子掲示版における会話からのハイライト部分の抽出，人工知能基礎論研究会（2002).

[2] 熊田 研治，久保 田稔: Twitter を用いたニュースストピックにおける少数意見の抽出，情報科学技術フォーラム，Vol.28, pp.1–4 (2015).

[3] 鳥海 不二夫，松澤 有，久保 田稔: ヴォーカルマイノリティ現象を説明する意見発信モデルの提案，情報処理学会論文誌，Vol.58, pp.1277–1286 (2017).