Vol.2018-ICS-192 No.6
2018/7/6

大規模ニュース記事データベースを用いた複数観点からの記事の対比と報道スタンスの分析

立浪 紀彦[1,a)]　吉岡 真治[1]

概要：多様な配信元によるインターネット上のニュース記事について、それぞれの記事が、取り上げているトピックに対してどのような姿勢（スタンス）で報道されているのかについて分析するために、統計的有意かつ比較対象とする記事群での支持度（アイテムの出現頻度/全データ数）の比が大きなパターンに注目した Statistical Emerging Pattern Mining を用いた分析手法を提案する。また、予備的な実験をおこない、その考察を行い、提案手法がトピックに特徴的な語を見つけるという性質を持つことを確認したが、ニュースサイトの分析という意味では、まだ不十分であることを確認した。今後は、ニュースサイトの特徴を、より、抽出しやすくするための比較対象とする記事群の設定方法などについて、考察を進めていくことが必要であることが分かった。

1. はじめに

近年、世界中の様々なニュースサイトから、大量のニュースが発信されている。また、これらのニュースサイトのニュースは、Yahoo News や Google News などのニュースアグリゲーションサイトを通じて、メジャーなサイトの記事から、マイナーな記事のサイトまで、多くの人に閲覧される可能性がある。一方、昔ながらの特定のニュースサイトにいって、ニュースを読むユーザについては、ニュースサイトの報道スタンス（右寄り、左寄り）を、理解しながら記事を読むことが想定されるが、ニュースアグリゲーションサイトなどを通じて、1 つだけの記事を取り出して読む場合には、このようなサイトの特徴を考慮して読むことは難しい。本研究では、各々のニュースサイトが様々な問題に対し、どのように報道するのかを分析することにより報道スタンスを分析する方法の提案を最終目標とする。

本論文では、まず、この問題を考えるにあたり、我々がこれまでに行ってきたニュースサイトの国別比較に関する研究 [1] や、特定の賛否に注目したニュースサイトの特性分析 [3] について紹介するとともに、本研究でその分析に利用する統計的有意性を保証するパターンマイニングの手法 [2] を紹介する。

さらに、この手法を用いて、世界中のニュース記事を集める大規模なデータベー [*1] が収集している日本語記事を対象とした分析実験を行い、その結果について考察を行った。

[1] 北海道大学
hokkaido Uniersity
[a)] s02140684b@eis.hokudai.ac.jp
[*1] https://www.gdeltproject.org/

2. ニュースサイトの比較分析

2.1 複数ニュースサイトの比較

我々は、これまでに、国別や地域別の複数のニュースサイト群を比較することにより、それらのサイトが属する国の興味の比較を行うシステムである NSContrast の提案を行ってきた [1]。この研究では、国別や地域別で集めたニュース群を対照比較することにより、これらの国や地域でのみ比較的多く報道されている事象を見つけ、各々の国や地域の興味を分析する手法であった。この研究では、ニュースサイトの対照比較をするという点では共通点があるが、ニュースサイトのグループ化には、地域性のみを考慮しているのみであり、本研究で目標としている報道スタンスによるニュースサイトの分類という考え方とは少し異なる。

我々は、また、ニュースサイトが扱う特定のトピックにおける賛否の割合の違いに注目し、報道スタンスの違いを可視化する方法を提案している [3]。この手法では、2.2 節で説明する GDELT のニュース記事に対する賛否の情報を用いて、特定のトピック（大統領選挙の時期のドナルド・トランプとヒラリー・クリントンに関する報道）における賛否の割合の違いに注目することにより、どちらのトピックにも同じような報道スタンスのニュースサイト、候補者毎の報道スタンスが大きく異なるサイトなどの存在を可視化することが可能となった。しかし、そのような報道スタンスが異なるサイトがどのような興味を持っているのかと

いった情報についての分析は不十分であった。

2.2 GDELT

上記のニュースサイトの比較分析では、ニュースサイトの情報として、世界中のニュースサイトからニュースを収集し、人物名や、賛否といったメタデータを付与して整理したデータベースである Global Database of Events,Language and Tone(GDELT) を利用している。この GDELT には、GDELT Event Database と GDELT Global Knowledge Graph(GKG) という 2 つのデータベースを提供している。GDELT Event Database は世界中で起こったイベントついてのデータベースであり、報道されているイベントについて ID を付与し、イベントごとに発生した日時、そのイベントについてはじめて報じられた日時、同じイベントについて報じている記事の数や Conflict and Mediation Event Observations(CAMEO) コードによって設定されたイベントの種類や関連する人物・組織をはじめ、イベントを報じている記事については記事の ID や長さ、賛否（トーン）の平均値、記事が元々書かれていた言語などの情報が登録されている。Global Knowledge Graph は記事についてのデータベースであり、記事が報じているイベント、関連する人物、組織名、-100 から+100 で表される賛否（トーン）の度合い、記事の URL が登録されている。GDELT では分析に使う記事は英語で書かれているもののほかに日本語を含む 65 の言語の記事を収集しており、それらを英語に機械翻訳して分析している。各種のデータには公式サイトや GoogleBigQuery などを通じて取得することが可能である。

3. 統計的有意性を保証するパターンマイニング

3.1 Stastical Emerging Pattern Mining(SEPM)

ニュースサイトの分析を行う際に、記事中に存在する語をアイテム、記事をパターンとして考えると、支持度（出現頻度/全データ数）に基づくパターンマイニングは、ニュースサイトの分析に有用な一手法であると考えられる。しかし、多くの場合、特定のトピックに関連する特徴語（例えば、首相の名前と首相）は、様々なニュースサイトで支持度が高い場合が多く、単純な出現頻度に基づくマイニングでは、本研究が対象とするニュースサイトの特徴を十分にとらえることは難しい。そこで、本研究では、対象とするニュースサイトにおける出現頻度が他のサイトの出現頻度と比べて特徴的に異なるパターン（記事中の出現語群）に注目した分析を考える。

このようなパターンを分析する方法として、2 つのデータセットにおいて、片方のデータセットにおいて支持度が高く、もう片方のデータセットでは支持度が低いパターンをその比によって特徴づける emerging Pattern を検出する Emerging Pattern Mining(EPM) がある。しかし、emerging pattern のマイニングは、支持度の比を使うため、支持度が低い方のデータセットの支持度が小さい場合に、その値が統計的な有意性を保証できない場合があり、このような統計的に有意でないパターンを結果に含めない Stastical Emerging Pattern Mining(SEPM)[2] の研究がある。本研究では、SEPM のアルゴリズムとして、次節で説明する多重比較検定の考え方を用いて統計的有意性を検証する QT-LAMP-EP を用いて分析を行う。

3.2 QT-LAMP-EP

統計的仮設検定における多重比較とは 3 つ以上の群に対して、どの群の間に有意な差があるか検定することであり、パターン同士を検定することでそのパターンに含まれるアイテムが特徴的なアイテムと考える。この際検定を行う数が多くなるため、検定の過誤が発生する確率が高くなる。ここで、全ての検定の内少なくとも 1 回以上誤って帰無仮説を棄却する確率を FWER(Family-Wise Error Rate) といい、誤って帰無仮説を棄却しない回数の期待値を FDR(False Discovery Rate) といい、多重比較検定では FWER と FDR を制御する必要がある。FWER と FDR の制御の方法である Bonferroni 補正では、設定した有意水準を仮説検定を行う数で割った値を各検定の有意水準とするが、パターンマイニングにおける多重仮説検定では検定の数が多すぎるために、各検定の有意水準が低くなりすぎてしまい、有意なパターンを見つけることが困難になってしまう。そのため FWER と FDR を制御しつつ有意水準を下げすぎないように補正する必要がある。

EPM における FWER と FDR の制御する手法である QT-LAMP EP[2] は多重仮説検定の際、検定可能性 (quasi-testability) に注目し、FWER と FDR に影響しないパターンを取り除き実際に行う検定の数を減らすことで、より多くのパターン発見する手法である。QT-LAMP-EP ではまずデータセット D を D_{main} と D_{carib} の 2 つに分割し、D_{carib} から検定を行うパターンと補正後の有意水準を計算し、それらを用いて D_{main} から統計的に有意なパターンを検出する。

4. 提案手法

本研究では、GDELT により収集された様々なニュースサイトから集めた大規模ニュース記事データベースを利用したニュースサイトの報道スタンスの分析を行うことを目標とする。具体的には、各ニュースサイトごとに、トピック毎に特徴的なキーワードの違いや、賛否の違いなどに関する情報を獲得することによって、ニュースサイトの報道スタンスを分析するための基本情報を提供することを想定している。

また、GDELT のメタデータには、本文情報が存在しないため、単語レベルでの分析ができなかった。その問題を

解決するために、本研究では、GDELT のデータに含まれている URL の情報からニュース記事のページにアクセスし、本文を抽出したうえで、形態素解析を行い、記事中の単語情報を獲得し、GDELT のメタデータの拡充を行う。一つのニュース記事に対応する GDELT のメタデータと記事中の単語情報を一つのアイテムセットとしたニュース記事のデータベースを作成し、分析を行う。

この特徴的なキーワードの獲得するために、特定の条件（ニュースサイト、トピックを表すキーワード、賛否、あるいは、それらの組み合わせ）を満たす記事群から作ったデータセットと、それ以外のデータセットを作成し、前節で述べた QT-LAMP-EP を適用し、特徴的なパターンの発見を行う。この時、特徴的なパターンは、与えた特定の条件を特徴づけるパターンとなる。

ただ、比較対象とする、それ以外のデータセットをどのように作るのが良いのかについては、検討の余地がある。例えば、ニュースサイトで制約をかけたデータセットと、それ以外のデータセットを比較することにより、そのニュースサイトが特徴的に興味を持つと考えられるキーワード群を得ることが可能となる。また、そのパターンの中に、単語だけのセットが現れる場合には、ニュースサイトが特徴的に興味を持つトピックと考えられ、賛否の情報と組み合わせて出てくる場合には、賛否も考慮した特徴的なトピックを表すと考えられる。

しかし、複数の制約条件を含むような詳しい制約条件を与えたときに、比較対象とする記事群の選択を、どのようにするのが良いのかを考える必要がある。例えば、ニュースサイトと賛成という条件で制約条件を与える場合には、それ以外の記事と比較すると、比較対象とする記事に、同じニュースサイトの中立や反対の記事を含むため、ニュースサイトの特徴が出にくくなる可能性がある。

本稿では、上記の問題を考慮し、簡単な制約条件についてのみの実験を行うこととした。

4.1 ニュース記事データベースの作成

GDELT が提供する記事のメタデータを取得する。記事本文については公開からしばらくすると元の Web ページから削除、または非公開状態になってしまう場合があるため、メタデータの取得時に得られた記事の URL を参照することによって保存した。この URL を参照した際に得られるページは HTML などで書かれており分析に不要なタグなどが含まれているため、データセットの作成の際には Web スクレイピングソフトである boilerpipe[4] を用いて本文を抽出する。

取得した記事には、以下のような理由で本研究に利用するには不適切なものが含まれている。本文を抽出した段階で得られたものを確認し該当のものサイトによる記事については本研究では利用しないこととした。

- 記事本文の閲覧が有料であり、直接 URL を参照するだけでは本文を取得できなかったもの
- 本文の抽出がうまくできなかったもの
- 記事の内容が個人のブログや商品の紹介などだったもの

4.2 統計的優位性を保証するパターンマイニング

取得したデータを用いてパターンマイニングを行う。ここでは、以下のデータを用いる。

- 記事の日付
- 記事に関する関連人物
- 記事を公開したサイト名
- 賛否の値（0,1,2）
- 記事本文に現れた単語

賛否の値には-100 から+100 の値が取得できるが、+100 から+1 を肯定的、+1 から-1 を中立的、-1 から-100 を否定的とし、3 値に分割した。記事本文に現れた単語については、今回利用した記事から全文検索を行い、全記事中 3 %以上の記事で現れたものの中から、記事のスタンスを分析する手がかりとなりそうなものを選んだ。一つの記事より一つのトランザクションが作られ、それぞれのトランザクションには日付、サイト名、賛否の値が 1 つずつ、さらに関連人物と記事本文に現れた単語が含まれる。

4.2.1 アイテムの表記ゆれへの処理

GDELT より取得した関連人物の項目と同じ人物を指す単語が本文中に現れている場合、ほぼ同じ意味のアイテムが複数存在してしまうため、これらはひとつのアイテムとして扱うことにした。

4.3 トピックによるデータセットの分割

emerging pattern を考えるにあたって比較する 2 つのデータセットを用意する必要があるが、本研究では、その第一段階として、トピック

報道スタンスを分析した際に特徴が現れるようなトピックを定め、それをアイテムとして含むか否かで 2 つのデータセットを作成した。

5. 実験結果と分析

5.1 記事の取得

GDELT より記事の元々の言語が日本語の記事で、2017 年 9 月 1 日から 2018 年 5 月 15 日までの記事を用いた。記事数は GDELT から取得した全記事数が 778,975、本文抽出の際のサイトの選定し予備実験に用いられる記事数は 119,523 となった。

記事が配信された時期については一ヶ月ごとに区切り、アイテムとして加えた。

5.2 データセットの作成

パターンマイニングを行うためのデータセットを作成した。QT-LAMP-EP を行うにあたり用意する 2 つのデータセットについては記事に関連するトピックとして「安倍晋三」を設定し、これに関連する記事群とそうでない記事群データセットを作成した。このとき d-他セット間で記事数が大きく変わらないように一部のデータを使い、「安倍晋三」を含む記事群は仮説検定の有意水準は 0.05、$a = 0.65$、QT-LAMP-EP を行う上でデータセットの 20% を D_{carib} とした。

5.3 結果

実行した結果、統計的に有意といえるパターンの内で、安倍晋三を含む記事群での支持度とその他の記事での支持度の比が大きいものの上位は次の通りである。「否定的」は、記事が否定的であることを示している。

- 選挙、北朝鮮
- 「否定的」、消費、選挙
- 東京、消費、選挙
- 日本、消費、選挙
- 「否定的」、消費、北朝鮮
- 2017 年 9 月、選挙、北朝鮮

これらのキーワードは、選挙と関係して、消費（消費税）や北朝鮮への対応が注目されていることを示している。しかし、これらの結果は、安倍晋三の特徴としては、有意味であるが、ニュースサイトの分析という観点からは、あまり、有用ではない。

次に、最上位ではないが、上位に含まれるパターンのうち、報道スタンスの分析の手がかりとなると考えられるようなニュースサイトを含むパターンを選ぶと、下記のようなものが見つかったが、全体としての総数は少数であった。

- 中立的、reuters.com、選挙、金融、米国、北朝鮮
- 2017 年 10 月 、中立的、tokyo-np.co.jp、選挙
- reuters.com 米国 選挙 北朝鮮

サイト名として出てくるのは、reuters のような外国の報道機関の日本語版といった一般の日本のニュースサイトと傾向が違うと考えられるサイトの「米国」との関係を含むようなパターンのみであった。ニュースサイトには、それなりの報道スタンスの特徴があると思うが、今回のように、一つのニュースサイトとそれ以外で統計的有意な差異が存在することが望まれるために、似たような傾向を持つサイトが複数存在するような場合には、特定のニュースサイトがパターンに含まれにくくなっているのではないか、と考えている。

6. おわりに

本論文では、GDELT によるニュース記事のメタデータと本文の情報に対して統計的有意性を保証するパターンマイニングを行うことでニュース記事群の際に注目した分析を行う方法を提案した。しかし、現時点の結果では、比較対象とする記事群の特徴をパターンとして得ることはできることを確認したものの、ニュースサイトの報道スタンスを分析するような情報が、十分に得られているわけではない。今後は、比較対象とする記事群をうまく設定することで、ニュースサイトの報道スタンスを分析する方法について、検討していきたい。

また、本研究では元々日本語で書かれた記事を対象としたが、英語版でも同様の分析実験を行いたいと考えている。

謝辞 本研究の一部は、JSPS 科研費 18H03338 の助成と北海道大学国際連携研究教育局ビッグデータ・サイバーセキュリティグローバルステーションの支援を受けた。ここに記して謝意をあらわす。

参考文献

[1] 吉岡真治，神門典子：複数国の新聞からの多観点比較による分析〜GDELT データを用いた分析〜，インタラクティブ情報アクセスと可視化マイニング研究会第 8 回研究会研究発表予稿集 (2014).
SIG-AM-08-05.

[2] Komiyama, J., Ishihata, M., Arimura, H., Nishibayashi, T. and Minato, S.-i.: Statistical Emerging Pattern Mining with Multiple Testing Correction, *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, New York, NY, USA, ACM, pp. 897–906 (2017).

[3] Yoshioka, M., Jang, M., Allan, J. and Kando, N.: Visualizing Polarity-based Stances of News Websites, *Proceedings of the Second International Workshop on Recent Trends in News Information Retrieval co-located with 40th European Conference on Information Retrieval (ECIR 2018)* (2018).
http://ceur-ws.org/Vol-2079/paper2.pdf.

[4] Kohlschütter, C., Fankhauser, P. and Nejdl, W.: Boilerplate Detection Using Shallow Text Features, *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, New York, NY, USA, ACM, pp. 441–450 (2010).