

オンラインニュースに対する Twitter 上での 地域局所的な注目の分析及び予測

長城 沙樹[†] 山口 祐人^{††} 北川 博之^{†††} 天笠 俊之^{†††}

[†] 筑波大学情報学群情報科学類 〒 305-8577 茨城県つくば市天王台 1-1-1

^{††} 筑波大学計算科学研究センター 〒 305-8577 茨城県つくば市天王台 1-1-1

^{†††} 筑波大学システム情報系情報工学科 〒 305-8577 茨城県つくば市天王台 1-1-1

E-mail: [†]s.nagaki@kde.cs.tsukuba.ac.jp, ^{††}yuto_ymgc@kde.cs.tsukuba.ac.jp

^{†††}{kitagawa, amagasa}@cs.tsukuba.ac.jp

あらまし 本研究では、オンラインニュースが Twitter 上で地理的にどのように注目されるかを分析・予測する。まず、特定地域の Twitter ユーザに言及されるニュースをローカルニュースと定義し、その性質をニュース記事のカテゴリ、本文、及び言及ツイートの地理的な分布という観点から分析する。また、ニュース記事本文及びカテゴリを用いてそのニュースが地理局所的に注目をあつめるか否かを予測する手法を提案する。

キーワード オンラインニュース, Twitter, マイクロブログ

1. 序 論

オンライン上で大量に発行されているニュースの中には、地域局所的に注目されるニュースと全国的に注目されるニュースが存在する。両者にはどのような違いがあるのだろうか。本研究では、Twitter^(注1)上で地域局所的に注目されるニュースがどのような特徴をもっているかを分析する。また、ニュース記事の本文及び記事のカテゴリを用いて、ニュース記事が地域局所的に注目されるか否かを予測できるかを検証する。本研究では、Twitter 上で話題になっていないニュースや、発行されて間もないニュースなど、Twitter 上で言及するユーザが少ないニュースについても、そのニュースが一部地域で注目されるか否かを予測することができることを明らかにした。各地域で局所的に注目されているニュースがわかると、あるニュースが局所的に注目されている地域に居住するユーザに、そのニュースを推薦することができる。逆に、全国的に注目されるニュースは全ユーザに推薦することができる。本研究によって、地域局所的に注目されるニュースと全国的に注目されるニュースの分類が可能になるため、ニュースの推薦に役立てることができると考えられる。

ニュースの記事内容を用いてそれらが注目される地域を評価するのは難しい問題である。この理由として、テレビや新聞などの様々なメディアの影響で、オンラインニュースの中には、内容が似たニュースであっても全国的に注目されるニュースや一部地域で注目されるニュースが存在することがあげられる。また、同じ内容のニュースであっても時間が経過するにつれて各ニュースが注目される地域は変化すると考えられる。本研究は、地域局所的に注目されるニュースと全国的に注目されるニュースでは、記事本文及びカテゴリに違いがあることを分析によって示し、それらを利用したナイーブベイズ分類器を作成するこ

とで、ニュース記事の地域局所性を判定した。

本研究では、Twitter ユーザ全体の地理的分布と Yahoo!ニュース^(注2)の URL が記載されたツイートを投稿したユーザの地理的分布を比較し、分布が大きく異なるニュースをローカルニュースと定義する。これは、Twitter を利用しているユーザ数が都道府県によって偏りがあり、単純にニュースの URL が記載されたツイートを投稿したユーザ数を地域ごとに比較することでは、ニュースの地域局所性を判定できないためである。ローカルニュースとローカルでないニュースにどのような特徴があるのかを、ニュース記事のカテゴリ、ニュース記事本文、ニュース記事の URL が記載されたツイート数、ニュースが発行された時刻からツイートが投稿されるまでの経過時間、ツイートを投稿した Twitter ユーザの地理的分布という観点から分析する。また、ナイーブベイズ分類器を用いたニュース記事の地域局所性判定手法を提案する。提案手法は、記事本文を入力すると、そのニュースがローカルニュースか否かを分類する。本研究では実際のニュースデータを用いて、提案手法の分類性能を検証する。本研究の貢献は以下の通りである。

- Twitter において地理局所的に注目されるニュースをローカルニュースとして定義し、その特徴として以下の点があることを示す。
 - － ローカルニュースとローカルでないニュースを比較すると記事のカテゴリの割合が異なる。
 - － ローカルニュースとローカルでないニュースに出現する名詞には違いがある。
 - － ローカルニュースと比較して、ローカルでないニュースは各ニュースの URL が記載されたツイート数のニュースによる差が大きい。
 - － ローカルニュースとローカルでないニュースでは、ニュー

(注1): <https://twitter.com/>

(注2): <http://news.yahoo.co.jp/>

スの URL が記載されたツイートを投稿するユーザの地理的分布が時間経過によって変化の様子が異なる。

- ニュース記事本文及びカテゴリを用いてその記事がローカルニュースであるか否かをある程度の精度を持って分類できることを示す。

提案手法によってニュース記事の本文及びカテゴリを用いてローカルニュースか否かを分類した結果、平均正解率は最大で 0.728 となった。提案手法によりニュースを分類した結果は、地名がタイトルに含まれているか否かを利用してニュースを分類した結果と比較して、正解率が 9 ポイント高かった。

本論文の構成は以下のとおりである。第 2 章では本研究の関連研究について述べる。第 3 章で本研究に用いたデータセットについて述べ、第 4 章では本研究におけるローカルニュースの定義と、定義したローカルニュースの特徴を示す。第 5 章で本研究で提案するニュース記事の地域局所性判定手法と、その評価実験について述べる。最後に第 6 章で本論文のまとめと今後の課題について述べる。

2. 関連研究

オンラインニュースを閲覧するユーザの地理的関係に関する研究として、Martin ら [1] の研究があげられる。Martin らは、Yahoo!ニュースで配信された記事とそれに対するコメントデータを用いて、地理的な距離が近いユーザ同士の方が、地理的に距離が遠いユーザ同士よりも同じ記事に対してコメントを行っていることを示した。また、Martin らは、アメリカの州レベルにおいて、犯罪率や失業率などの社会経済的な指標やその地域に居住する人々の性格と、その地域に居住するユーザが読む記事のトピックには関連があることも示した。Martin らは、Yahoo!ニュースを対象にして研究を行ったが、Gebrekirstos ら [2] は、複数のニュースポータルサイトを対象として、ニュースの記事を閲覧するユーザの地理的関係を分析した。その結果、総合的なポータルサイトの「地域カテゴリ」に属す記事は、その地域のユーザから閲覧される割合が高いことを示した。また、同研究において Gebrekirstos らは、総合的なニュースポータルサイトのニュース記事は、あるカテゴリに特化したポータルサイトのニュース記事と比較して、ニュース記事を閲覧するユーザの地域に偏りがあることを示した。

オンラインニュースに限らず、オンラインサービスが提供する内容とサービスを利用するユーザの地理的関係についての研究が行われてきた。中でも Anders ら [3] は、YouTube に投稿された動画の視聴データを用いて、動画が視聴される地域には動画によって偏りがあること、また、動画の視聴数が最も多い日から時間が経過するにつれて、動画が視聴される地域は限られてくることを示した。さらに Anders らは、外部サイトからのリンクや直接 URL を入力することによって動画が視聴された割合が高いほど、地域局所的に動画が視聴されていることを示した。

本研究ではニュースの地域性を測るために、Twitter を利用した。Twitter は近年多くのユーザに利用されており、そのユーザに関する様々な研究が行われてきた。Kwak ら [4] は、Twitter ユーザのフォロー関係を分析し、Twitter をソーシャル・ネット

ワーキング・サービスとしてだけでなく、情報収集のためにも使っているユーザが多いことを示した。また、Twitter で互いにフォローしているユーザ同士は地理的に近い割合が高いことも発見した。Culotta ら [5] は、Twitter のフォロー関係と Web サイトの閲覧情報から Twitter ユーザの属性を推定した。この結果、ごく少数のフォロワー情報からそのユーザの人種や性別、年齢などを推定することができることを示した。

Twitter を情報収集の手段として使用するユーザが多い [4] という背景から、オンラインニュースとソーシャルメディアでの話題を比較した研究が行われてきた。Alexandra ら [6] は、オンラインサービス上で話題となっている出来事を抽出する手法を提案した。さらに提案手法を用いて、気候変動に関して、ニュースメディア上で話題となる出来事とソーシャルメディア上で話題となる出来事を抽出し、それらを比較した。その結果、ニュースメディアで報道される出来事とソーシャルメディアで共有される出来事には違いがあることを示した。Leetaru ら [7] は、ニュースサイトで配信されている記事に関連する地域の分布とツイートの発信地の分布を世界規模で比較し、それらが大きく異なるということを示した。Hua ら [8] は、複数のデータセットでの相互作用を考慮したトピックモデルを作成し、Twitter とニュースメディアで話題となるトピックを比較した。その結果、同じ出来事に対しても Twitter とニュースメディアでは注目する点が異なること、一般的に Twitter で話題となるトピックよりもニュースメディアで話題となるトピックの方が、他メディアで取り上げられやすいことを示した。

本研究は、ソーシャルメディアにおいて一部地域で注目されるニュースを定義し、ニュース記事本文及びカテゴリを用いて各ニュースの地域局所性を判定するという点で関連研究と異なる。

3. データセット

本研究では、Twitter に投稿されるニュースとして、Yahoo!ニュースを用いた。Twitter Streaming API^(注3)を用いて、2015 年 11 月 24 日から 2015 年 12 月 23 日の間に、Yahoo!ニュースのドメイン ‘headlines.yahoo.co.jp’ を含む URL が記載されたツイートを投稿しているツイートデータ 1,517,056 件及び、ツイートされたページの HTML データ 84,789 件を収集した。収集した HTML データから動画ニュースのデータや別サイトのデータを取り除いた。残りの HTML データから記事のタイトル、本文、カテゴリ、発行日時を抽出し、使用するニュースデータとした。作成したニュースデータは 76,275 件となった。

次に、収集したツイートデータについて、投稿したユーザの居住地を収集した。収集したツイートデータ 1,517,056 件を投稿したユーザは 444,648 アカウントだった。OpenStreetMap の Nominatim^(注4)を用いて、ユーザ自らがテキストデータで設定した居住地情報から都道府県を抽出し、ユーザの居住する都道府県とした。444,648 のユーザアカウントを用いて 130,097 ユーザの居住する都道府県を収集した。残りのユーザは、居住

(注3): <http://dev.twitter.com/streaming/overview>

(注4): <http://wiki.openstreetmap.org/wiki/Nominatim>



図 1 ニュースの URL が記載されたツイートを投稿したユーザ数を都道府県ごとに色分けした図。カラーバーは各都道府県に居住するユーザ数のカラースケールを表し、色が濃くなるほどユーザ数が多いことを示す。東京都に居住する Twitter ユーザ数は非常に多く、都道府県によって Twitter を利用しているユーザ数に偏りがあることがわかる。

地を設定していないか、地名以外のテキストデータを居住地として設定しているユーザであった。収集したユーザ数を Google Chart API の GeoChart^(注5) を用いて、都道府県ごとに色分けした結果を図 1 に示す。カラーバーは各都道府県に居住するユーザ数のカラースケールを表し、色が濃くなるほどユーザ数が多いことを示す。図 1 から、東京都に居住する Twitter ユーザ数は非常に多く、都道府県によって Twitter を利用しているユーザ数に偏りがあることがわかる。

4. ローカルニュース

本研究では、Twitter ユーザの地理的分布とニュースの URL が記載されたツイートを投稿したユーザの地理的分布を比較し、分布が大きく異なるニュースをローカルニュースとする。4.1 節で本研究におけるローカルニュースの定義を示す。4.2 節では、実際のニュースデータ、ツイートデータを分析し、4.1 節で定義したローカルニュースの特徴を示す。

4.1 ローカルニュースの定義

地域局所的に注目されるニュースを定義する方法として、そのニュースの URL が記載されたツイートを投稿したユーザ数を各地域で比較し、ある地域から投稿したユーザ数が一定数以上のニュースをローカルニュースと定義する方法が考えられる。しかし、図 1 に示された通り、Twitter を利用しているユーザ数は都道府県によって偏りが大きい。そのため、ニュースの URL が記載されたツイートを投稿したユーザ数を比較すると、Twitter を利用しているユーザ数が多い都道府県のユーザ数が多くなると考えられる。そこで、各都道府県を確率変数とし、各都道府県に居住するユーザ数の割合を確率値とする離散確率分布を、ユーザの地理的分布と定義する。Twitter ユーザ全体の地理的分布とニュースの URL が記載されたツイートを投稿したユーザの地理的分布を比較し、分布が大きく異なるニュースを

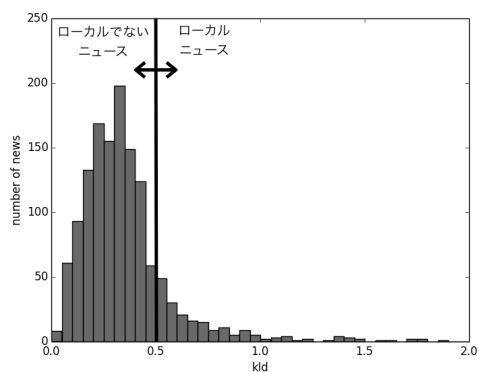


図 2 すべてのユーザの地理的分布からみた、ニュースの URL が記載されたツイートを投稿したユーザの地理的分布の KL ダイバージェンスのヒストグラム。ビンの幅は 0.05 であり、x 軸が KL ダイバージェンスの値、y 軸がニュース数を示す。ニュースの多くは KL ダイバージェンスの値が 0.0~0.5 に集中し、右に裾がない分布となっていることがわかる。

ローカルニュースとする。

Twitter ユーザ全体の地理的分布とあるニュースの URL が記載されたツイートを投稿したユーザの地理的分布の比較には、Kullback-Leibler(KL) ダイバージェンスを用いる。KL ダイバージェンスは、2 つの確率分布の差異を図る尺度であり、確率分布 Q からみた確率分布 P の KL ダイバージェンスは式 1 で計算される。

$$KL(P \parallel Q) = \sum_x P(X=x) \log \frac{P(X=x)}{Q(X=x)} \quad (1)$$

KL ダイバージェンスを用いることで、都道府県ごとのユーザ数の偏りを考慮したうえで、あるニュースについて注目するユーザに地理的な偏りがあるのかを示すことができる。

本研究では、以下のようにローカルニュースを定義する。

定義 (ローカルニュース)

ニュースの URL が記載されたツイートを投稿したユーザの地理的分布を P とする。また、すべてのユーザの地理的分布を Q とする。あるニュースの URL が記載されたツイートのうち、居住する都道府県が分かっているユーザが投稿したツイートが 50 件以上あるとき、 Q からみた P の KL ダイバージェンスが 0.5 以上となるニュースをローカルニュースとする。

第 3 章で述べたデータセットを用いたところ、居住地がわかっているユーザからの投稿が 50 件以上あるニュースは、76,275 件中 1,348 件であった。これらの KL ダイバージェンスのヒストグラムを図 2 に示す。ビンの幅は 0.05 であり、x 軸が KL ダイバージェンスの値、y 軸がニュース数を示す。ニュースの多くは KL ダイバージェンスの値が 0.0~0.5 に集中し、右に裾がない分布となっていることがわかる。KL ダイバージェンスを計算した結果、ローカルニュース (local) は 199 件、ローカルでないニュース (non-local) は 1,149 件となった。

(注5): <https://developers.google.com/chart/interactive/docs/gallery/geochart>

4.2 ローカルニュースの特徴

本節では 4.1 節で定義したローカルニュースの特徴を示す。まず、4.2.1 節と、4.2.2 節で、ローカルニュースとローカルでないニュースの記事にどのような違いがあるかを示す。4.2.1 節では Yahoo!ニュースのカテゴリに着目し、4.2.2 節では本文の違いに着目する。さらに残りの節で、ニュースの URL が記載されたツイートにどのような違いがあるかを示す。4.2.3 節で、ニュースの URL が記載されたツイートの数に着目し、4.2.4 節では、それらが時間経過によってどのように変化するかを示す。最後に、4.2.5 節では、ニュースの URL が記載されたツイートを投稿したユーザの地理的分布が時間によってどのように変化するかを示す。

4.2.1 カテゴリの分布

Yahoo!ニュースのニュース記事は、あらかじめそのニュースがどのような分野の記事であるか、8 つのカテゴリ（国内、国際、経済、エンタメ、スポーツ、IT・科学、ライフ、地域）に分けられている。本節では、ローカルニュースとローカルでないニュースの内容の違いを、全ニュースに占める Yahoo!ニュースの各カテゴリに属する記事の割合によって示す。

ローカルニュース、ローカルでないニュースに占める、それぞれのカテゴリに属する記事の割合を計算した。結果を図 3 に示す。x 軸が各カテゴリの記事が占める割合、y 軸が Yahoo!ニュースのカテゴリを示す。結果から、ローカルでないニュースと比較して、ローカルニュースの中には、「スポーツ」カテゴリや「地域」カテゴリの記事の割合が高いことがわかる。記事の内容を確認めると、同じ「スポーツ」カテゴリの中でも、ローカルニュースの記事にはプロ野球チームに関する記事が多く、ローカルでないニュースの記事には野球以外のスポーツに関する記事が多かった。「地域」カテゴリのニュースには、他メディアなどで全国的に報道されるようなニュースは少ない。そのため、「地域」カテゴリのニュースは、各出来事が起きた地域で注目されやすいのではないかと考えられる。逆に「国際」カテゴリの記事が日本国内において地域局所的には注目されにくいのも、同様の理由によると考えられる。すなわち、各出来事が起きた地域と関係するユーザは少なく、他メディアや Twitter 上でニュースを知るユーザが多いからだと考えられる。また、「エンタメ」カテゴリのニュースもローカルでないニュースに多いことがわかる。テレビや映画は全国的に放映されているものが多く、全国から注目されていると考えられる。

4.2.2 自己相互情報量による特徴語抽出

本節では、ローカルニュースとローカルでないニュースの記事本文の違いがあることを示すため、ローカルニュース、ローカルでないニュースの特徴語を抽出し、その結果を示す。特徴語の抽出には、自己相互情報量 (pointwise mutual information, PMI) を用いる。自己相互情報量は、確率変数 X, Y とそれぞれの事象 x, y について、2 つの事象 x, y の間の共起の度合いを測る尺度の 1 つであり、式 2 で定義される。

$$PMI(X = x, Y = y) = \log \frac{P(X = x, Y = y)}{P(X = x)P(Y = y)} \quad (2)$$

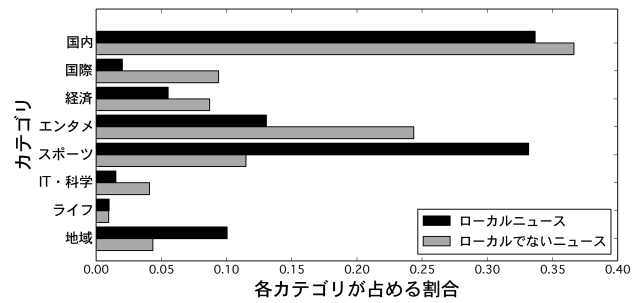


図 3 ローカルニュース・ローカルでないニュースの全記事に占める、Yahoo!ニュースの各カテゴリに属する記事の割合。x 軸が各カテゴリの記事が占める割合を示す。y 軸が Yahoo!ニュースのカテゴリを示す。ローカルでないニュースと比較して、ローカルニュースには「スポーツ」カテゴリや「地域」カテゴリのニュースが多いことがわかる。ローカルでないニュースには「国際」カテゴリや「エンタメ」カテゴリのニュースが多いことがわかる。

表 1 自己相互情報量上位 10 件の名詞を降順に示す。

ローカルニュース		ローカルでないニュース	
PMI	名詞	PMI	名詞
1.987	移籍	0.206	消費
1.880	広島	0.203	視聴
1.760	橋下	0.202	警視庁
1.686	プレー	0.198	背景
1.679	来季	0.198	フジテレビ
1.638	福岡	0.198	カメラ
1.564	クラブ	0.196	主演
1.551	決勝	0.196	テーマ
1.516	市長	0.196	人間
1.487	リーグ	0.193	作品

分析の手法は以下の通りである。まず、ニュース記事を用いて bag-of-words を作成する。ニュース記事に出てきた名詞の中から、全記事で出現数が 30 以下、もしくは 3 割以上の記事に出現する名詞を省き、bag-of-words とする。bag-of-words の作成には、MeCab^(注6) 及び Python ライブラリ gensim^(注7) を用いる。得られた名詞の集合を X 、ニュースの種類 (ローカルニュース: local, ローカルニュースでないニュース: non-local) を Y とし、自己相互情報量を計算する。 $Y=local$ の時に自己相互情報量が大きくなる語を、ローカルニュースの特徴語と考えることができる。ニュース記事データ 1,348 件の中から、最終的に得られた名詞は 877 語となった。得られた各名詞と $Y=local, non-local$ の自己相互情報量を計算した。自己相互情報量上位 10 件の名詞を表 1 に示す。

$Y=local$ としたとき、自己相互情報量が大きい単語には、プロ野球に関する語や、市政・県政に関する語が多く含まれていた。 $Y=non-local$ としたとき、自己相互情報量が大きい単語には、「消費」「視聴」「作品」など、一部の地域から注目されるとは考えにくい単語が多く見られる。この結果から、プロ野球や市政・県政に関するニュースは、Twitter 上で地域局所的に注

(注 6): <http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html>

(注 7): <https://radimrehurek.com/gensim/>

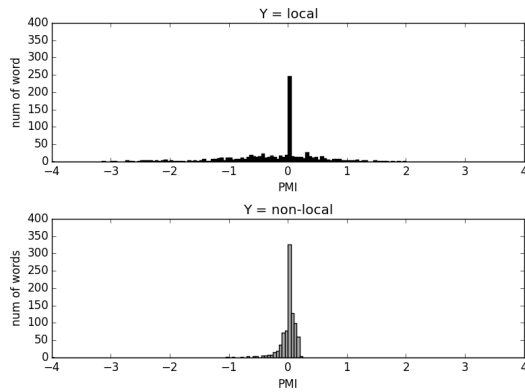


図 4 各名詞とローカルニュース（上）・ローカルでないニュース（下）の自己相互情報量のヒストグラム。ビンの幅は 0.05 で、x 軸が自己相互情報量、y 軸が各名詞の数を表す。ローカルニュースに比べてローカルでないニュースは自己相互情報量が 0 近辺に集中していることがわかる。

目されやすいといえる。4.2.1 節で、ローカルニュースに占める「スポーツ」カテゴリの記事の割合が大きいことを示した。プロ野球に関するニュースは「スポーツ」カテゴリに分類されるため、本節の結果は、4.2.1 節の結果とつながっているといえる。また、ユーザが居住している地域に直接関係する市政・県政に関するニュースは、地域局所的に注目されやすいといえる。

表 1 から、ローカルニュースとローカルでないニュースで、自己相互情報量上位 10 件の値が大きく異なることがわかる。求めた自己相互情報量のヒストグラムを、図 4 に示す。ビンの幅は 0.05 で、x 軸が自己相互情報量、y 軸が各名詞の数を表す。図 4 から、ローカルニュースに比べてローカルでないニュースは自己相互情報量が 0 近辺に集中していることがわかる。ローカルでないニュースは、ローカルニュースとくらべて記事の数が多く、内容也多岐にわたっていると考えられる。そのため、特定の語がローカルでないニュースと共起しやすいとは考えにくい。従って、ローカルでないニュースの自己相互情報量の絶対値は小さくなるのではないと思われる。

4.2.3 ニュースの URL が記載されたツイート数の差

本節では、ローカルニュースとローカルでないニュースで、ニュースの URL が記載されたツイート数の違いを示す。

ローカルニュース・ローカルでないニュースの URL が記載されたツイート数の累積確率分布 (cdf) を図 5 に示す。x 軸がツイート数（対数スケール）、y 軸が累積確率分布を示す。図 5 から、ローカルでないニュースはローカルニュースと比較して、あるひとつのニュースの URL が記載されたツイート数が多い傾向があり、加えてニュースごとの差が大きいことがわかる。ローカルニュースに分類されるニュースには、ある特定の地域で起きた出来事に関するニュースが多く含まれると推測される。そのため、ニュース記事に関する地域に居住するユーザはそのニュースに注目しやすく、それ以外の地域に居住するユーザはあまり注目しないと考えられる。逆に、ローカルでないニュースは、テレビや新聞など他メディアの影響が大きいため、注目するユーザが多いと考えられる。そのためローカルニュースと比

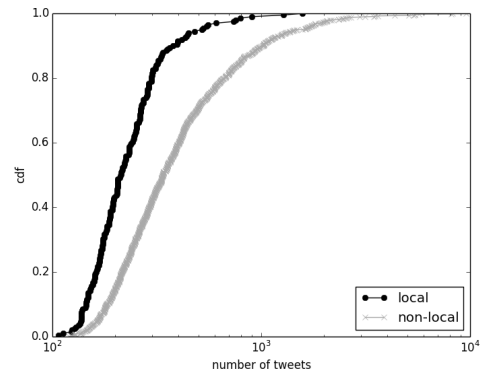


図 5 ローカルニュース（黒）・ローカルでないニュース（灰色）の URL が記載されたツイート数の累積確率分布。x 軸がツイート数（対数スケール）、y 軸が累積確率分布を示す。ローカルでないニュースはローカルニュースと比較して、あるひとつのニュースの URL が記載されたツイート数が多い傾向があり、加えてニュースごとの差が大きいことがわかる。

較して、各ニュースに関するツイート数も多く、また非常に多くのユーザから関心をもたれるニュースが存在するのではないかと考えられる。

4.2.4 時間経過によるツイート数の変化

本節では、あるニュースが Twitter 上で拡散される様子がローカルニュースとローカルでないニュースで異なるかを示す。4.2.3 節において、ニュースの URL が記載されたツイート数は異なることが示された。そこで、あるニュースの URL が記載された全ツイートに占める、そのニュースが発行された時刻からの各経過時間までに投稿されたツイート数の割合を求め、経過時間によってどのように変化するかを示す。

まず、各ツイートについて、そのツイートに URL が記載されたニュースの発行時刻とツイートが投稿された時刻との差を計算する。次に各ニュースについて、ニュースの URL が記載された全ツイート数に占める、各経過時間までに投稿されたツイート数の割合を計算する。最後にローカルニュース、ローカルでないニュースにおける平均値を計算する。結果を図 6 に示す。x 軸がニュースが発行された時刻からツイートが投稿されるまでの経過時間、y 軸が各ニュースの URL が記載された全ツイート数に占める、各時間までに投稿されたツイート数の割合を示す。エラーバーは標準偏差を示す。

4.2.3 節で述べた通り、ローカルニュースはローカルでないニュースと比べて各ニュースに言及されたツイート数は少ない。しかしながら図 6 より、ローカルニュースとローカルでないニュースで、各ニュースに関するツイート数に占める、各時間までに投稿されたツイート数の割合の差は小さいことがわかる。また、ローカルニュースもローカルでないニュースも、ニュースが投稿されてから時間が経過するにつれてツイートは少なくなっていることがわかる。Twitter では、他ユーザが投稿したツイートを引用して投稿 (ReTweet, RT) することで、他ユーザが投稿したツイートの内容を自分のフォロワーに共有することができる。Twitter ユーザは他のユーザとの情報共有のために

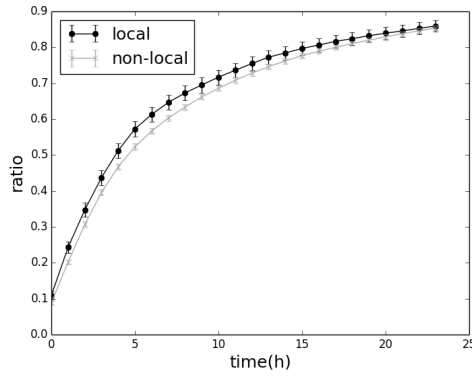


図 6 各ニュースの URL が記載された全ツイート数に占める、ニュースの発行時刻からの各経過時間までに投稿されたツイート数の割合の平均値。x 軸がニュースが発行された時刻からツイートが投稿されるまでの経過時間、y 軸が各ニュースの URL が記載された全ツイート数に占める、各経過時間までに投稿されたツイート数の割合を示す。エラーバーは標準偏差を示す。ローカルニュースとローカルでないニュースで、各ニュースに関するツイート数に占める、各時間までにツイートされたツイート数の割合の差は小さいことがわかる。ローカルニュースとローカルでないニュースのどちらも、ニュースが投稿されてから時間が経過するにつれてツイートは少なくなっていることがわかる。

RT を利用することが多い。先行研究により RT のうち 50%は、元のツイートが投稿されてから 1 時間以内になされており、元のツイートが投稿されてから時間が経過するにつれて RT される確率は小さくなることが示されている [4]。そのため、ニュースが投稿されてから時間が経過するにつれてツイートは少なくなっているのではないかと考えられる。このように、ローカルニュースとローカルでないニュースでは、Twitter 上で拡散される様子にほとんど違いはないといえる。

4.2.5 時間経過によるユーザの地理的分布の変化

ニュースが発行されてからの経過時間による、各ニュースの URL を記載したツイートを投稿したユーザの地理的分布の変化について示す。まず、各ツイートの URL が記載されたニュースの発行時刻とツイートが投稿された時刻との差を計算する。次に各ニュースについて、各経過時間 t までに投稿されたツイートを行ったユーザの地理的分布 P_t を作成する。さらに、全ユーザの地理的分布を Q とし、 Q からみた P_t の KL ダイバージェンスを計算する。最後に全ニュースについて、計算した KL ダイバージェンスの平均値を計算する。結果を図 7 に示す。x 軸がニュースの発行時刻からの経過時間 t を示す。y 軸が Q からみた P_t の KL ダイバージェンスを示す。エラーバーは標準偏差を示す。

ローカルニュースとローカルでないニュースのいずれも、ニュースが発行されてから時間が経過するにつれて KL ダイバージェンスの値が小さくなっていることがわかる。これは、ニュースが発行されてから時間が経過するにつれて、地理的に広い範囲に居住するユーザにニュースが拡散されるという直感と一致する。また、ローカルニュースもローカルでないニュースも、ニュースが発行されてから時間が経過するにつれて、KL ダ

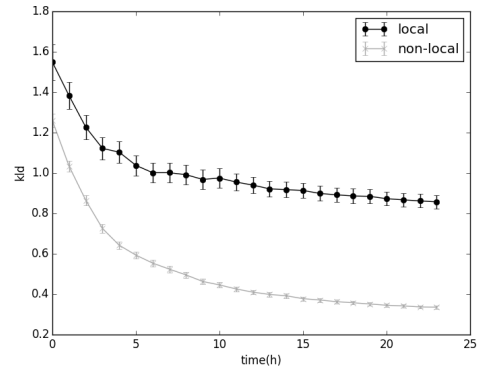


図 7 ニュースの発行時刻から各経過時間までに、ニュースの URL が記載されたツイートを投稿したユーザの地理的分布の変化。x 軸がニュースの発行時刻からの経過時間 t を示す。y 軸が全ユーザの地理的分布からみた、各経過時間 t までにツイートを投稿したユーザの地理的分布の KL ダイバージェンスを示す。エラーバーは標準偏差を示す。ローカルニュースとローカルでないニュースのいずれも、ニュースの発行時刻から時間が経過するにつれて、KL ダイバージェンスの値が小さくなっており、その変化も小さくなっている。さらに、ローカルニュースの方がニュースが発行されてから早い段階で変化が小さくなることからわかる。

イバージェンスの値の変化は小さくなっている。これは、4.2.4 節で述べたように、ニュースが発行されてから時間が経過するにつれて、ニュースの URL が記載されたツイートが投稿されにくくなっているからではないかと考えられる。さらに、ローカルニュースの方がニュースが発行されてから早い段階で変化が小さくなることからわかる。これは、ローカルニュースは一定の地域以外に居住するユーザに広まりにくいと想定される。

5. ニュースの分類

第 4 章において、本研究で定義したローカルニュースとローカルでないニュースには、ニュース記事の本文の内容と記事のカテゴリの割合に違いがあることを示した。本章ではナイーブベイズ分類を利用し、ニュース記事本文及びカテゴリを用いてそのニュースがローカルニュースかローカルでないニュースかに分類できるか検証した結果を示す。5.1 節で本研究における提案手法について示し、5.2 節で実際に行った評価実験について述べる。

5.1 提案手法

本研究で提案するニュース記事の地域局所性判定手法には、ナイーブベイズ分類を利用する。ナイーブベイズ分類は式 3 で表されるベイズの定理を利用した分類法である。

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \quad (3)$$

提案手法では、式 3 から導出される式 4 を利用する。

$$\begin{aligned} P(loc|doc, cat) &= \frac{P(loc)P(doc, cat|loc)}{P(doc, cat)} \\ &\propto P(loc)P(doc, cat|loc) \\ &= P(loc)P(doc|loc)P(cat|loc) \end{aligned} \quad (4)$$

ここで、 $P(loc|doc, cat)$ は、カテゴリ cat である文書 doc が与えられた時に、その文書が local(non-local) である確率を示す。 $P(cat)$ は総文書中に占める各カテゴリの文書の割合を示す。 $P(doc|loc)$ は、ローカルニュース（ローカルでないニュース）が与えられた時に文書 doc が生成される確率を示す。 $P(cat|loc)$ は、ローカルニュース（ローカルでないニュース）が与えられた時にその文書がカテゴリ cat である確率を示す。

文書 doc に出現する単語の集合を $\{word_1, word_2, \dots, word_k\}$ として表し、これが文書 doc を示すと考える。各単語が出現する確率は独立であると仮定すると、式 5 が成立する。

$$\begin{aligned} P(doc|cat) &= P(word_1 \wedge word_2 \wedge \dots \wedge word_k|cat) \\ &= \prod_i P(word_i|cat) \end{aligned} \quad (5)$$

$T(cat, word)$ をカテゴリ cat の記事に単語 $word$ が出てきた回数、 V を訓練データ中の全単語集合とすると、 $P(word_i|cat)$ は、式 6 で表すことができる。

$$P(word_i|cat) = \frac{T(cat, word_i)}{\sum_{word' \in V} T(cat, word')} \quad (6)$$

しかし未知の文書のカテゴリを推定する際、訓練データに全く含まれない単語が 1 つでも含まれると、 $P(word|cat)$ が 0 となるため、式 5 より、 $P(doc|cat)$ が 0 となる。それを防ぐためにパラメータ α を式 7 のように導入する。

$$P(word_i|cat) = \frac{T(cat, word_i) + \alpha}{\sum_{word' \in V} T(cat, word') + |V|\alpha} \quad (7)$$

(ただし、 $\alpha \geq 0$)

式 4～式 7 から式 8 が導出される。

$$\begin{aligned} &P(loc|doc, cat) \\ &= P(loc) \left\{ \prod_i \frac{T(cat, word_i) + \alpha}{\sum_{word' \in V} T(cat, word') + |V|\alpha} \right\} P(cat|loc) \end{aligned} \quad (8)$$

最後に、文書 doc とカテゴリ cat の重みを調整するパラメータ β を式 9 のように導入する。 $P(local|doc, cat)$ をパラメータ β で調整した結果を local-point、 $P(non-local|doc, cat)$ をパラメータ β で調整した結果を non-local-point とする。

$$\begin{aligned} &local-point \\ &= P(loc) \left\{ \prod_i \frac{T(cat, word_i) + \alpha}{\sum_{word' \in V} T(cat, word') + |V|\alpha} \right\}^\beta P(cat|loc)^{(1-\beta)} \end{aligned} \quad (9)$$

提案手法では、式 9 を用いて、local-point 及び non-local-point を計算する。それらと比較して、カテゴリ cat である文書 doc がローカルニュースか否かを予測する。

5.2 評価実験

本節では、提案手法で実際のニュースデータ 1,348 件を分類できるか検証した結果を示す。5.2.1 節にて実験方法、5.2.2 節にて実験結果を示す。

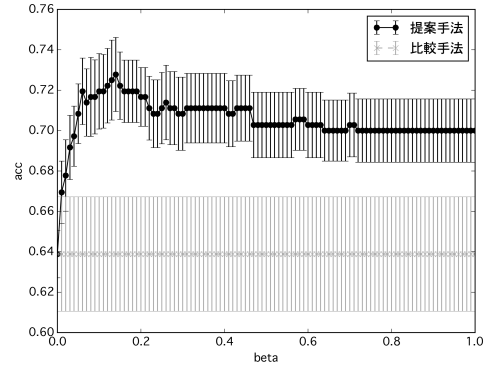


図 8 評価実験の結果. x 軸がパラメータ β の値, y 軸が正解率の平均値を示す。パラメータ β が 0.14 の時、提案手法の精度は最も良く、比較手法よりも 9 ポイント高かった。

5.2.1 実験方法

4.2.2 節で作成した bag-of-words に含まれる名詞の集合を用いて、あるニュースがローカルニュースか否かを判断するナイーブベイズ分類器を作成する。ナイーブベイズ分類器の作成は、Python ライブラリの scikit-learn^(注8) から、MultinomialNB を利用する。パラメータ α は 1.0 とする。また、パラメータ β を 0 から 1 まで 0.01 ずつ変化させる。精度を検証するため、データを 10 分割して交差検定を行い、正解率の平均値を計算する。正解率を式 10 で定義する。 $N_{a,b}$ は、正解ラベル a 、予測したラベル b であるニュース数とし、 N_{all} を全ニュース数とする。

$$acc = \frac{N_{local,local} + N_{non-local,non-local}}{N_{all}} \quad (10)$$

提案手法と比較するために、タイトルに地名が入っているニュースをローカルニュース、タイトルに地名が入っていないニュースをローカルでないニュースと分類し、正解率を計算する。地名の抽出には MeCab を用いる。地名の有無を本文でなくタイトルに限った理由は、ほとんどのニュース記事には地名が含まれているからである。提案手法の実験と同様のテストデータを分類し、正解率の平均値を計算する。

5.2.2 実験結果

評価実験の結果を図 8 に示す。x 軸がパラメータ β の値, y 軸が正解率の平均値を示す。提案手法で最も良い精度となったのは、パラメータ β が 0.14 の時であり、そのときの正解率の平均値は 0.728、標準偏差 0.0183 となった。比較手法の正解率の平均値は 0.639、標準偏差は 0.0283 となった。提案手法による分類結果は、比較手法による分類結果よりも 9 ポイント精度が高かった。タイトルに地名が入っているニュースの中には、他メディアなどの影響で全国的に有名になっているニュースも存在する。提案手法で分類した結果を確認すると、そのようなニュースをローカルでないニュースとして正しく分類できていた。

提案手法によって、ローカルニュースに誤って分類されたローカルでないニュースのうち、ランダムに抽出した 10 件を表 2 に、ローカルでないニュースに誤って分類されたローカルニュース

(注 8): <http://scikit-learn.org/stable/>

表 2 ローカルニュースに誤って分類されたローカルでないニュース (一部抽出)。

カテゴリ	タイトル
スポーツ	300点超え優勝の羽生「いまだに信じられない」
スポーツ	C大阪、フィッカデンティ監督浮上 すでに水面下で接触済み
国内	忘年会で最終電車乗り過ごし、救済バス...高尾発
スポーツ	オリックス中島が改名を発表 来季から「裕之」→「宏之」に
地域	御堂筋をフェラーリのF1カーが疾走 「大阪でF1レースをやるべき」 橋下市長 / 大阪
スポーツ	【クラブW杯】ネイマール快気祝い!?表参道ふらっと爆買い2000万円
スポーツ	球児、涙の会見「自分の人生の中心に阪神タイガース」「全身全霊いつ倒れてもいい」
エンタメ	橋下徹氏、古巣タイタンに復帰 タレント活動へ
スポーツ	なぜ長居?問い合わせ殺到でJリーグが"4位C大阪ホーム"のPO決勝の経緯を説明
スポーツ	西武・森 2年目野手球団最高も...球団が「歩き方」にダメ出し

表 3 ローカルでないニュースに誤って分類されたローカルニュース (一部抽出)。

カテゴリ	タイトル
ライフ	外猫を捕まえて持ち込み 保健所は引き取る? 弁護士が解説
国内	民主が野党統一候補を検討 参院選1人区共闘呼びかけへ
エンタメ	倅田来未、初の47都道府県ツアー「全国をリアルに回ろうと思います」
国際	拉致再調査合意発表から29日で1年半 被害者家族「早く帰って新たな生活できるように」
国内	三重・伊勢25.4度の「夏日」 各地で12月最高気温
国内	<福岡・天神の通り>「親富孝」派?「親不孝」派?街の50人に聞く
地域	歴史的価値ある「幻の新橋駅」、活用方法探る 東京メトロ
エンタメ	東MA X、桃食べて死にそうに...アナフィラキシーショックの怖さ語る
エンタメ	EXILE x SHIBUYA109 が Xmas コラボキャンペーン実施
国内	<京都市>路上喫煙は過料1000円 徴収される外国人急増

のうち、ランダムに抽出した10件を表3に示す。提案手法によって、ローカルニュースに誤って分類されたローカルでないニュースには、「スポーツ」カテゴリの記事が多く存在した。これはローカルニュースの中に、スポーツに関する記事が多く含まれているためだと考えられる。逆に提案手法によって、ローカルでないニュースに誤って分類されたローカルニュースには、芸能界に関するニュースや国内で起きた出来事に関するニュースが多くあった。芸能界に関するニュースには、メディアの影響で全国的に注目されるニュースもあれば、地理局所的に注目されるニュースもある。後者の例としては、ある芸能人がある地域で行ったイベントに関するニュースが挙げられる。しかし、提案手法では、この2つを分類することができなかった。また同様に、国内で起きた出来事に関するニュースにも、全国的に注目されるニュースと、地域局所的に注目されるニュースが存在する。提案手法ではこれらを正確に分類することができなかった。

6. ま と め

本研究では、Twitterにおいて地理局所的に注目されるニュースをKLダイバージェンスを用いて定義し、どのような特徴をもっているか分析した。分析の結果、(1) ローカルニュースとローカルでないニュースを比較すると記事のカテゴリの割合が異なること、(2) ローカルニュースとローカルでないニュースに出現する名詞には違いがあること、(3) ローカルニュースと比較して、ローカルでないニュースは各ニュースのURLが記載

されたツイート数のニュースによる差が大きい、(4) ローカルニュースとローカルでないニュースでは、ニュースのURLが記載されたツイートを投稿するユーザの地理的分布が時間経過によって変化する様子が異なることが示された。また本研究は、ニュース記事の本文及びカテゴリを用いたナイーブベイズ分類器によるニュース記事の地理的局所性を分類する手法を提案し、提案手法は本研究で定義したローカルニュースを正解率0.728で予測できることを示した。本研究によって、Twitter上で話題になっていないニュースや、発行されて間もないニュースなど、Twitter上で言及するユーザが少ないニュースについても、そのニュースがローカルニュースか否かを予測することができる。今後の課題としては、分析で得られたローカルニュースの特徴を活用し、さらに精度の高い予測手法の検討が挙げられる。具体的には、ローカルニュース・ローカルでないニュースの特徴語を利用した予測手法が考えられる。また、推薦に利用できるまでにある一定量のツイートデータが必要となるが、ニュースが発行してある程度の時間が経過するまでに、ニュースのURLが記載されたツイートを投稿したユーザの地理的分布も利用できると考えられる。さらに、本研究で提案した各ニュースの地域局所性を予測する手法を用いて、各ニュースが局所的に注目される特定地域を予測し、ニュースの推薦手法に応用したい。

謝 辞

本研究の一部は、文部科学省「実社会ビッグデータ利活用のためのデータ統合・解析技術の研究開発」による。

文 献

[1] S.Martin, Q.Daniele, M.Amin, “The Geography of Online News Engagement” Proceedings of the 6th international conference on Social Informatics, pp.279-289, November 2014.

[2] G.Gebrekirstos, P.Arjen, “The Role of Geographic Information in News Consumption” Proceedings of the 24th International Conference on World Wide Web, pp.755-760, May 2015.

[3] B.Anders, S.Salvatore, W.Mirjam, “YouTube Around the World: Geographic Popularity of Videos” Proceedings of the 21st international conference on World Wide Web, pp.241-250, April 2012.

[4] H.Kwak, C.Lee, H.Park, S.Moon, “What is Twitter, a social network or a news media?” Proceedings of the 19th international conference on World Wide Web, pp.591-600, April 2010.

[5] A.Culotta, NK.Ravi, J.Cutler, “Predicting the Demographics of Twitter Users from Website Traffic Data” Proceedings of the 29th International AAAI Conference on Artificial Intelligence, pp.72-78, January 2015.

[6] O.Alexandra, C.Carlos, D.Nicholas, A.Karl, “Comparing Events Coverage in Online News and Social Media: The Case of Climate Change” Proceedings of the 9th International AAAI Conference on Web and Social Media, pp.288-297, May 2015.

[7] KH.Leetaru, S.Wang, G.Cao, A.Padmanabhan, E.Shook. “Mapping the global Twitter heartbeat: The geography of Twitter” First Monday, Volume 18, Number 5 - 6 May 2013.

[8] T.Hua, Y.Ning, F.Chen, CT.Lu, N.Ramakrishnan. “Topical Analysis of Interactions Between News and Social Media” Proceedings of the 30th AAAI Conference on Artificial Intelligence, Feb 2016.(to appear)