6K-04

Twitter上の社会・地理・内容的特徴を用いたローカルニュースの抽出

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　†筑波大学計算科学研究センター
‡筑波大学システム情報系情報工学域

1　序論

オンラインニュースには, 一部地域に居住するユーザから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, その特徴を示す. また, ニュース記事の内容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. 本研究を応用することで, ニュース記事を用いてそのニュースが注目される地域を推測し, ユーザにあったニュースを推薦することができると考えられる. 提案手法によってニュース記事の本文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0.7となった. 提案手法により, 地名がタイトルに含まれているかでローカルニュースか否かを判断した結果よりも, 正解率が約6ポイント向上した.

2　関連研究

オンラインサービスとユーザの地理的関係について, Twitter上のフォロー関係とユーザの居住地の関連性の研究 [1] や, オンラインニュース記事と記事を読むユーザが居住する地域との関連性の研究 [2] など様々な研究がなされている. また, オンラインニュース上で話題となるトピックとTwitter上で話題となるトピックを比較し, 違いがあることが示されている [3].

本研究は, ニュース記事の内容からそのニュースがTwitterにおいて地理局所的に注目されるか予測するという点で関連研究と異なる.

Local News Extraction in Twitter Based on Social, Geographical, and Textual Characteristics
Saki NAGAKI§, Yuto YAMAGUCHI†,
Hiroyuki KITAGAWA†‡and Toshiyuki AMAGASA†‡
§College of Information Science, University of Tsukuba
†Center for Computational Sciences, University of Tsukuba
‡Faculty of Engineering, Information and Systems, University of Tsukuba

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を用いて, 2015年11月24日から2015年12月23日の間に投稿された, 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1,517,056件及び, 投稿されたページのHTML84,789件を収集した. 収集したHTML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き, 記事のタイトル, 記事の本文を含むニュースデータ76,275件を作成した.

次に, Twitterユーザ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. 収集したツイートデータを投稿したユーザは444,648アカウントであり, そのうち30,097ユーザの居住する都道府県を収集できた.

4　ローカルニュース

本章では, 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.

4.1　ローカルニュースの定義

Twitterを利用しているユーザの数は都道府県の偏りが大きい. そのため本研究では, 2つの確率分布の差異を図るKLダイバージェンスを利用し, ローカルニュースを以下のように定義する.

ローカルニュース 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件以上あるニュースのうち, すべてのユーザが居住する都道府県の確率分布からみた, ニュースのURLをツイートしたユーザが居住する都道府県の確率分布のKLダイバージェンスが0.5以上となるニュース

第3章のデータを用いて, ローカルニュース199件, ローカルでないニュース1,149件を作成した.

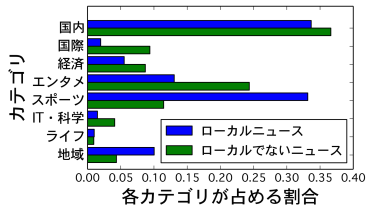

図 1: ローカルニュース（青）・ローカルでないニュース（緑）のうち, 各カテゴリに属す記事の割合を示す.

表 1: PMI 上位 5 件

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

4.2 ローカルニュースの特徴

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違いを 2 つの観点から示す.

カテゴリの分布　Yahoo!ニュースの記事は, 8 カテゴリに分けられる. ローカルニュース, ローカルでないニュースのうち, 各カテゴリの割合を図 1 に示す.

ローカルニュースに占める「国内」カテゴリの記事が多いのは, ある出来事に関する記事は, その発生地に居住するユーザの興味をひきやすいためではないかと推測できる. また, ローカルニュースに「スポーツ」カテゴリの記事が多いのは, 各地域で活躍するプロ野球チームの影響によるのではないかと考えられる.

特徴語抽出　自己相互情報量 (PMI) は, 確率変数 X, Y とそれぞれの事象 x, y について, x, y の間の共起の度合いを測る尺度である. ニュース記事に出てきた名詞から, 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 を作成し, 各名詞とローカルニュース, ローカルでないニュースの PMI を計算する. PMI 上位 5 件の名詞を表 1 に示す.

ローカルニュースとの PMI が大きい単語には, 先に述べたようにプロ野球に関する語が多く含まれる. ローカルでないニュースとの PMI が大きい単語には, 娯楽に関する語が含まれている. これは, 一部地域にのみ関係する娯楽は少ないためと考えられる.

5 ニュースの分類

本章では, 実際のニュースデータをローカルニュースか否かに分類できるか検証した結果を示す.

実験方法　4.2 節で作成した bag-of-words を用いて, ナイーブベイズ分類器を作成する. パラメータ α は 1.0 とする. 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.

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.

実験結果　提案手法の正解率の平均値は 0.70, 標準偏差 0.049 となった. タイトルに地名が入っているか否かを利用した分類の正解率の平均値は 0.64, 標準偏差は 0.090 となった. 提案手法は, 地名の有無で分類するよりも精度が約 6 ポイント向上した. タイトルに地名が入っているニュースには, 他メディアなどの影響で全国的に有名なニュースも存在し, 提案手法はそのようなニュースを正しく分類できていると考えられる.

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. また, ニュース記事の本文を用いて, そのニュースがローカルニュースとなるか 0.70 の正解率で推測できることを示した. 今後は, 定義したローカルニュースの特徴を活用し, 分類手法の精度向上を目指す.

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.

参考文献

[1] H.Kwak, C.Lee, H.Park, S.Moon: What is Twitter, a social network or a news media? (WWW2010)

[2] S.Martin, Q.Daniele, M.Amin: The Geography of Online News Engagement (SocInfo2014)

[3] O.Alexandra, C.Carlos, D.Nicholas, A.Karl: Comparing Events Coverage in Online News and Social Media: The Case of Climate Change (ICWSM2015)