

閲覧実績を用いたニュース記事の地域性抽出

News Locality Detection using Users' Feedback Logs

大倉 俊平 *1

Shumpei Okura

*1 ヤフー株式会社

Yahoo Japan Corporation

News articles often deal with topics related to a specific location. Therefore, they are valuable when they are correctly delivered to users in the area, whereas worthless for users in other areas. In this paper, we propose a method to detect such locality of a news article. This method learns a recurrent neural network using users' click logs rather than human annotated data. Experimental result shows the proposed method achieves better detection accuracy than a filtering method based on a place name in its text.

1. はじめに

Web サービスにおいて、機械学習を用いてクリック傾向を学習し、過去の履歴から推定されるユーザの興味に応じて自動的にコンテンツを出し分けることが一般的になってきている [Covington 16]. また、スマートフォンの普及に伴って、GPS を用いてユーザの位置情報が取得できるようになり、それを元に現在地に近い店舗やイベントを推薦する推薦サービスも見られるようになった [Wanaka 16, Wang 15].

本稿では、ニュース記事に対して、そのニュースに興味を示す地域を推定する手法について述べる。例えば、特定地域からアクセスしてきたユーザに対して、その地域に強く紐づく記事を、新聞の地域面のような形で提供したり、興味情報から推薦した記事の中で、遠隔地に関する記事をフィルタするなどの応用が考えられる。

我々は、記事の地域性の推定において、記事配信に対するユーザの反応を利用する方法を提案する。閲覧実績データから、記事内容と反応するユーザの偏りの関係を学習する。この手法の利点は大きく2つある。

一つは、人手でアノテートされた学習データが不要であることである。例えば、記事と配信すべき地域のペアを学習データとして用意できれば、既存の教師あり学習のフレームで、地域性の推定器を構築することは可能である。しかし、各記事がローカルな話題であるかは、その地域の文化の理解が必要なのもあり、客観的な判断が難しい。それらを、各地域ごとに十分な量用意するのは、人的コストの面で現実的ではない。また、仮に一度作成したとしても、ニュース記事の時事的な性質を考えると定期的に新規の学習データを追加する必要があり、メンテナンスコストも高い。提案法では配信実績を利用するため、人手の作業なく大量の学習データが生成できる。また、サービス運用によって自動的に新規のサンプルが得られるため、新しい時事トピックへの適応や入稿される記事傾向の変化へ追従のための再学習も容易である。

もう一つは、記事に興味を示す地域を領域で捉えられることである。記事中に含まれる地名を参照して記事と地域を関連付ける手法を考える。例えば、記事中に「名古屋」という記述がある場合、地名と位置の辞書さえ事前に用意しておけば、

その記事は名古屋市周辺に配信すれば良いだろうことはわかる。しかし、その範囲が「名古屋市内」なのか「愛知県内」なのか「中部地方」であるのかは自明ではない。あるいは、そのニュースが非常にインパクトが大きく、全国のユーザが見るべきニュースである可能性もある。提案法では、反応したユーザの広がりから、その記事に興味を示す対象範囲も同時に得ることができる。

実験では、提案法を用いて特定地域に紐づく記事を抽出したところ、地名辞書を用いて抽出するベースラインと比べて、高い精度で抽出できることがわかった。

2. 提案手法

提案手法では、記事ごとの正解地域のラベルの代わりに、ユーザのクリック実績のデータを用いて記事と各地域の適合度を定義する。ただし、ニュース記事の場合は、入稿後すぐに正しい地域に配信できることが重要であるため、実績データをためてから配信に用いるのでは遅い。そこで、過去の配信記事を用いて記事内容から各地域の適合度を推定する推定器を学習しておき、新着記事の配信時には、その推定値をもとに地域性の有無を判定する。

2.1 地域適合度

記事と地域の適合度を実績データから定義する方法について述べる。まず、日本全体を緯度経度によって、 $N \times M$ のグリッドに分割する。次に、記事 x_k の配信結果をアクセス元の緯度経度情報を元に各グリッドに振り分け、グリッド (i, j) におけるクリック率 $r_k^{i,j}$ を算出する。 $r_k^{i,j}$ と、その平均値との乖離 $y_k^{i,j}$ を記事 x_k のグリッド (i, j) における適合度とみなす。具体的には、 $y_k^{i,j}$ は以下の式によって求める。

$$r_k^{i,j} = \frac{c_k^{i,j}}{v_k^{i,j}} \quad (1)$$

$$\bar{r}_k = \frac{1}{NM} \sum_{i,j} r_k^{i,j} \quad (2)$$

$$\sigma_k = \sqrt{\frac{1}{NM} \sum_{i,j} (r_k^{i,j} - \bar{r}_k)^2} \quad (3)$$

$$y_k^{i,j} = \frac{r_k^{i,j} - \bar{r}_k}{\sigma_k} \quad (4)$$

連絡先: 大倉俊平, ヤフー株式会社, 〒102-8282 東京都千代田区紀尾井町 1-3 東京ガーデンテラス紀尾井町 紀尾井タワー, sokura@yahoo-corp.jp

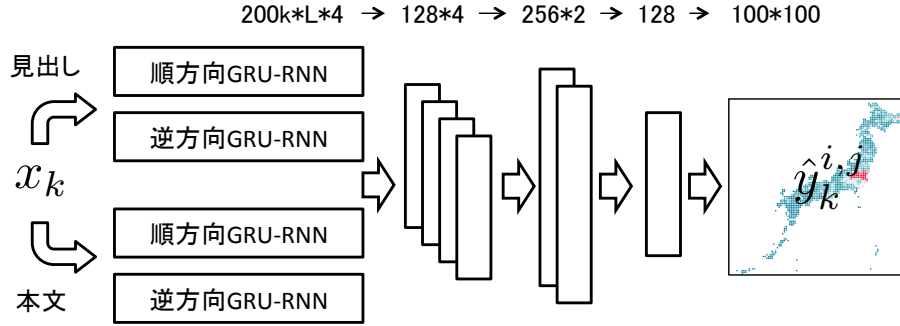


図 1: ニューラルネットワークの構成イメージ

ここで、 $v_k^{i,j}, c_k^{i,j}$ は、それぞれグリッド (i, j) に対する、記事 x_k の配信数とクリック数を表す。式 (4) 中の α は $[0, 1]$ の範囲の正規化パラメータである。 $\alpha = 1$ のときは通常の標準偏差での正規化であるが、その場合は実質的に地域差のない記事に対しても無理やり差を見出すことになる。一方で、 α が小さすぎると、配信数が少なくクリック率が安定しない場合の影響を強く受ける。本稿の実験では、 $\alpha = 0.5$ とした。

本手法において重要な点は、クリック率そのものではなく、他グリッドと比較した時の高低を用いることである。クリック率が高い記事でも、全国どこからでも読まれる記事の適合度は低くなり、逆にクリックが少なくても地域が偏っている記事は、特定グリッドの適合度が高く算出される。

2.2 推定モデル

次に、過去の配信実績データを用いて、記事 x_k から適合度 $y_k^{i,j}$ を推定する推定器を構築する。本手法では、人手のアノテーションにより学習データを作成する場合と比較して、大量の学習データが得られる一方で、それぞれのデータはノイズを多分に含んでいる。そのため、確率的勾配降下法により各サンプルを反復する学習法と相性が良い。

推定器は、比較的数の多い文書データを扱える手法ならば任意の手法を適用可能である。ここでは、図 1 に示す構造の recurrent neural network(RNN) を構築した。各記事は、見出しと本文をそれぞれ形態素に分割したのち、GRU[Cho 14] で構成された RNN を通して固定長のベクトルに変換する。その後、中間層を 3 層通して、推定値 $\hat{y}_k^{i,j}$ を出力する。全結合層の活性化には ReLU、最終層の損失関数は二乗損失 $\sum_{k,i,j} (\hat{y}_k^{i,j} - y_k^{i,j})^2$ を用いた。

3. 実験

前章で定義した地域適合度とその推定モデルがうまくはたらくことを示すため、実際のニュース記事と配信データを用いて検証を行った。

3.1 実験設定

推定モデルの学習には、Yahoo! JAPAN のスマートフォン版トップページにおける 2016 年 10 月～12 月のニュース配信結果を用いた。総配信数 $\sum_{i,j} v_k^{i,j}$ が 5000 以上の約 50 万記事に対して、その実績値から算出した適合度を、上記のネットワークに学習させた。形態素は出現頻度上位 199,998 語に未知語 (UNKNOWN) と文末 (EOS) を加えた 20 万語彙を使い、地域の分割は $100*100$ とした。学習はバッチサイズ 100 のミニバッチ SGD を用いて行い、学習率調整には RMS-

Prop[Tieleman 12] を利用した。適合度の正規化に関するパラメータ α は 0.5 とした。

3.2 評価

3 人の評価者に対して、それぞれの出身地に関する記事を以下の手順で抽出し、抽出された各記事がローカル記事としてふさわしいか否かを 2 択で判断してもらった。評価指標は各抽出手法で生成したリストにおけるローカル記事の含有率とする。評価対象の記事は 2017 年 2 月に入稿されたものを使用した。

- ランダムに 4 万件を抽出する (X)
- Xの中から、対象地域の県名・都市名を含む記事を 100 件抽出する (A)
- Xの中から、対象地域の推定値 $\hat{y}_k^{i,j}$ が高い順に 100 件抽出する (B)
- Bのうち、対象地域の県名・都市名を含む記事 (C)

評価結果は表 1 の通りである。3 人の評価者全てが、抽出方法 A よりも、抽出方法 B のほうがローカル記事として妥当なものを多く含むと判定した。つまり、上記の学習方法によって、単純な地名検索よりも精度の良いモデルを生成できたと言える。実用上は、両者を組み合わせた抽出法 C によってさらに効率の良い抽出が可能である。

抽出法 A の誤りは、複数地名が列挙されていたり、記事の本題と関係の薄いところで地名の記述がある記事を拾ってしまうものが多い。具体的な例としては、全国の統計値を比較している記事が全ての地域でローカル記事として抽出されてしまう場合や、記事内でコメントを引用している場合に発言者の出身地が付記されている場合などがあった。

提案法による地域性推定を用いた抽出法 B の場合、そのような記事はクリック率に地域差がつきにくく、同様の誤りが減っていた。一方で、スポーツや芸能関連の記事で、ローカル記事ではないが地域ごとのクリック率に差がつく記事が多く誤検出された。これらの記事は、地域ごとに最員のチームが違ったり、地元出身の著名人の場合に反応しやすかったりするため、クリック率の観点では地域差がつきやすい。しかし、それら全てがローカルな情報か否かと問われると、そうではないと感じるケースも多いようである。

抽出法 B と C を比較すると、僅かではあるが C には含まれていないローカル記事が存在することがわかる。これは、地名が明示的に記述されていないとしても、文脈から地域が限定される場合である。地名が含まれない記事について地域性判定が可能であることも、提案法の利点の一つであるといえる。

表 1: 評価結果

	評価者 1	評価者 2	評価者 3	平均
A	46/100 (46%)	35/100 (35%)	46/100 (46%)	42%
B	52/100 (52%)	46/100 (46%)	51/100 (51%)	50%
C	51/62 (82%)	37/44 (84%)	40/44 (91%)	86%

3.3 モデルに対する考察

ニューラルネットワークの動きを考察するため、最終層の $128 \rightarrow 100 * 100$ の部分の接続重みを可視化した。図 2 には、128 ノードのうち 30 ノードを掲載した。該当ノードが励起したときに反応するエリアを赤色で示した。

学習データには県境の情報は明示的に入っていないが、特定の都道府県の形状を表すノードが幾つか学習されていることがわかる。一方で、複数の県にまたがっている東北地方や近畿地方などに相当するとみられるノードも見つかった。これらの組み合わせにより、ピンポイントに特定地域だけに刺さる場合と、広い地域で興味を引く場合を表現できている。

4. まとめと今後の課題

本稿では、地域性の高いローカル記事の抽出方法の一つを提案した。提案法では、人手による正解データの作成なしで学習することができ、辞書マッチによる抽出より良い精度を達成した。人手による正解データの作成が不要なことは、話題の移り変わりの激しく、定期的な再学習が必要なニュース記事のドメインにおいて非常に有用である。本研究では、客観的な正解の定義が難しく本稿では 3 人の評価者の主観を含む評価を用いたが、複数の手法を客観的に評価できる評価指標を定めることは今後の課題のひとつである。今後は、作成したモデルを用いて実際に記事配信を行い、ユーザの反応を追っていく予定である。

参考文献

- [Cho 14] Cho, K., Merrienboer, van B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y.: Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (2014)
- [Covington 16] Covington, P., Adams, J., and Sargin, E.: Deep neural networks for youtube recommendations, in *Proceedings of the 10th ACM Conference on Recommender Systems*, pp. 191–198, New York, NY, USA (2016), ACM
- [Tieleman 12] Tieleman, T. and Hinton, G.: Lecture 6.5-RMSProp, COURSE: Neural networks for machine learning, *Technical report* (2012)
- [Wanaka 16] Wanaka, S. and Tsubouchi, K.: Location History Knows What You Like: Estimation of User Preference from Daily Location Movement, in *Proceedings of the Second International Conference on IoT in Urban Space, Urb-IoT '16*, pp. 8–13, New York, NY, USA (2016), ACM

- [Wang 15] Wang, W., Yin, H., Chen, L., Sun, Y., Sadiq, S., and Zhou, X.: Geo-sage: A geographical sparse additive generative model for spatial item recommendation, in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1255–1264 ACM (2015)



図 2: 学習されたモデルの最終層の接続重み (128 ノードのうち 30 を表示)