

ニュース記事の効率的な リアルタイムソーシャルアノテーション手法

大西 誠[†] 北川 博之[‡]

[†] 筑波大学情報学群情報科学類

[‡] 筑波大学システム情報系情報工学域

1 はじめに

ソーシャルメディアの普及に伴い、多くのユーザがインターネット上で情報収集、情報発信出来るようになった。中でも、Twitter^{*1}と呼ばれるマイクロブログでは、ツイートという140文字以内の短いメッセージを投稿することができ、ユーザが関心を持った出来事に対して、自分の意見や主張をテキストとしてリアルタイムで発信している。これらのツイートには、社会的に取り上げられているニュース記事に関する情報を含んでいることがある。その場合に投稿される内容の多くは、そのユーザが着目したニュースの補足的な内容であったり、重要部分を強調する内容であるため、そのニュースの特徴を知る上で非常に有益な情報であると言える。

近年、このようなツイートをリアルタイムで集める手法として Top-k Publish-Subscribe Approach [1] が提案された。これは、各ニュースとツイートの両方に出現する単語の数と、ツイートが投稿されてからの経過時間によって類似度を計算し、その中から類似度が高いツイートを k 個取得するものである。しかし、ニュースに関連するツイートの数はニュース毎に異なるため、全てのニュースに対し同一数のツイートを取得するこの手法は、あまり得策ではない。

我々の分析によれば、Twitter は即時性の高いマイクロブログであるため、ニュースが掲載される以前にも、そのニュース関連するツイートが発信されていることが多い。この特性を利用して、ニュースが掲載される以前のツイートとの類似度を計算し、それらのツイートの類似度の分布から各ニュースごとに適切な閾値を決定し、ニュース掲載後に投稿される関連ツイートのみを適切に選択することを試みる。本研究では、利用者が多く、データを公開しているという点から、Twitter と Yahoo!News^{*2}を用いる。

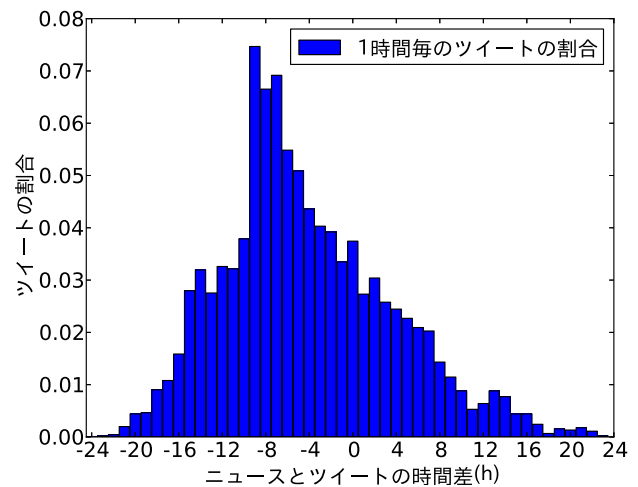


図1 各時間における関連ツイートの割合

2 関連ツイートの分布

本稿では、「ニュースに含まれている文章と同様の内容のツイート」と「ニュースに対して言及を行ったことが明らかなツイート」のことを関連ツイートと呼ぶ。図1は、本稿4節で述べている評価実験のデータにおいて、各ニュースに対する関連ツイートがいつ、どのくらい投稿されたのかを1時間毎の割合で表したものである。ニュースが掲載された時間を0とし、関連ツイートがそれより前に投稿された場合は負の時間、後に投稿された場合は正の時間で表現している。この図から、関連ツイートの多くはニュースが掲載されるより前に投稿されていることが分かる。

3 提案手法

どのニュースに対しても、類似度が低いツイート数は非常に多く、類似度が高いツイート数は非常に少ないという特徴がある。そのため、各ニュース毎にツイートとの類似度を計算し、その分布を調べ、ツイートの数が少なくなった所の類似度を閾値とすることで、そのニュースと関連性の高いツイートのみを得ることができる。閾値の決定にはニュースが掲載される前のツイートを使い、それ以降に投稿されるツイートは、閾値以上のものだけを取得する。

Efficient real-time social annotation of news articles

Sei Oonishi[†] (sei0024@kde.cs.tsukuba.ac.jp),

HiroYuki Kitagawa[‡] (kitagawa@cs.tsukuba.ac.jp)

[†] College of Information Science, University of Tsukuba

[‡] Division of Information Engineering, Faculty of Engineering, Information and Systems, University of Tsukuba

^{*1} <http://twitter.com/>

^{*2} <http://news.yahoo.co.jp/>

3.1 形態素解析

形態素解析によって類似度計算の対象となる単語を決定する．形態素解析には MeCab^{*3}を用いる．ストップワードのような特徴を持たない単語を取り除くために，単語は品詞が名詞で品詞細分類が一般，固有名詞，形容動詞語幹，サ変接続のものを処理の対象とする．

3.2 類似度計算

類似度の計算には Top-k pub-sub [1] における Content-based score とほぼ同様の式を用いる．ニュースを s ，ツイートに u とした場合，計算式は以下のようになる．

$$score(s, u) = \sum_i u_i * idf^2(i) * \sqrt{\frac{s_i}{|s|}} \quad (1)$$

s_i はニュース内における単語 i の出現回数， $|s|$ はニュース内の全単語数を表す．また，ニュース集合を S とした場合，逆文書頻度は $idf(i) = 1 + \log(|S| / (1 + |\{s \in S | s_i > 0\} |))$ とする．Content-based score との違いは， u_i がツイートにおける単語 i の出現回数ではなく，ツイート内で単語 i が出現した場合 1，出現しなかった場合は 0 の値をとる点である．これは，同じ単語を何度も含むような，意味の無いツイートの $score(s, u)$ を高くしないためである．もし，ニュース s とツイート u が共通の単語を含んでいれば， $score(s, u) > 0$ となる．

3.3 閾値の導出

新しくニュース s が掲載された時，それより前に保存しておいたツイート集合 U に対し式 (1) を用いて $score(s, u)$ の計算を行う．そのうち $score(s, u) > 0$ ($u \in U$) のツイートの集合を U_s とし，この集合 U_s における最大の $score(s, u)$ を $score_{max}$ とする．また，類似度を x としたとき， $d(x)$ を $x - \delta \leq score(s, u) < x + \delta$ ($u \in U_s$) であるツイートの個数を範囲 2δ で割ったものと定義する．このとき， $d_{max} = \max_{0 \leq x \leq score_{max}} d(x)$ とすると，類似度 x を $0 \leq x \leq score_{max}$ の範囲で昇順に調べ，以下の条件を満たす最初の x を閾値とする．

$$d(x) \geq d_{max} * p \quad (2)$$

p はユーザが与えるパラメータであり，これは，類似度の分布において，ツイートの数が最大の所から，どの程度減少した所で閾値を設定するかを決める値である．

4 評価実験

2013 年 12 月 20 日に投稿された 65 万個の日本語のツイートと，Yahoo!News のトップに掲載された 95 件のニュースを対象に評価実験を行った． $score(s, u)$ 中の idf を有効に利用するために，対象のニュースを含

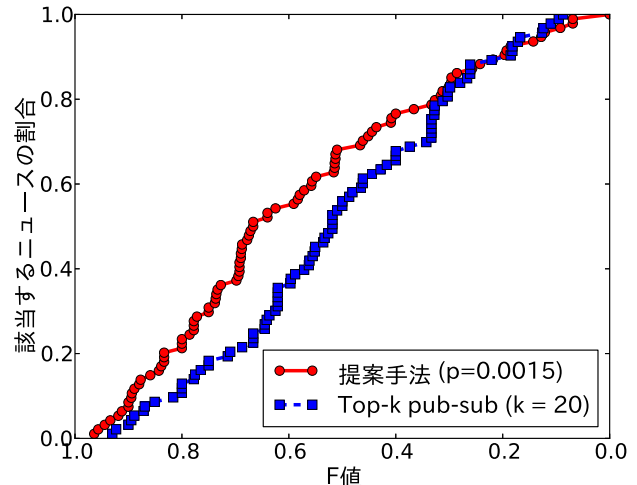


図2 F 値による比較評価

んだ約 300 件のニュースを用いて idf の計算を行っている．

これらのデータに対して，Top-k pub-sub と提案手法を用いた時の F 値の比較を行う．各パラメータはいくつかの値を比較して，全ニュースの F 値の平均が最大になるような値に設定した．今回の実験では $k = 20$ ， $\delta = 2.5$ ， $p = 0.0015$ とした．図 2 は横軸を F 値，縦軸はその F 値以上のニュースの割合を示したものである．提案手法は，Top-k pub-sub を用いた場合に比べ，F 値が高いニュースの数が多く，逆に F 値が低いニュースの数は少ないことが分かる．これは，ツイートの数が少ない類似度のところに関連ツイートが多く，ニュース毎に類似度の分布を用いて閾値を決定することが，全てのニュースに対して同一数のツイートを取得するよりも有用であることを示している．

5 まとめ

本稿では，ニュースが掲載される以前のツイートとの類似度を計算し，それらのツイートの類似度の分布から適切な閾値を決定する手法を提案した．さらに，この手法が，全てのニュースに対して同一数のツイートを取得するよりも有用であることを示した．今後の課題として，精度だけでなく処理速度の比較評価も行う予定である．

参考文献

- [1] Alexander Shraer, Maxim Gurevich, Marcus Fontoura, Vanja Josifovski, Top-k Publish-Subscribe for Social Annotation of News, Proceedings of the 39th International Conference on Very Large Data Bases, VLDB Endowment (2013).
- [2] Yahoo!JAPAN 媒体資料, http://i.yimg.jp/images/advertising/info/ad/pdf/Media_sheet_Open.pdf.

^{*3} <https://code.google.com/p/mecab/>