D-011

ニュース記事のリアルタイムソーシャルアノテーション
Real-time social annotation of news articles

大西誠 † Sei Onishi
北川博之 ‡ Hiroyuki Kitagawa

1. はじめに

マイクロブログの普及に伴い，多くのユーザが気軽に情報を収集・発信できるようになった．特に，Twitter 等のマイクロブログではツイートと呼ばれる短いメッセージにより，多種多彩な情報がユーザ同士で共有されている．Twitter ユーザが投稿するツイートの中には，ニュースの情報を含むツイートが多く存在する [1]．このようなニュースと関連のあるツイートは，ニュースの追加情報やニュースに関心を持つユーザの情報などを知ることができるため非常に有益な情報であると考えられる．

ニュースと関連のあるツイートを集める方法としてニュースとツイートの類似スコアを計算し，一定以上の類似スコアを持つツイートのみを集める方法が考えられる．しかし，ニュース毎に適切な類似スコアの閾値が異なるため，全てのニュースに対して一律の閾値を定めると，関連のあるツイートを集め損ねたり関連のないツイートを多く集めてしまう恐れがある．

本研究では，ニュース毎に適切な閾値を定め，ニュースが配信された時間より後に投稿された関連のあるツイートを集めることを目的とする．アプローチとして，ニュースが配信されるより過去に投稿されたツイートとの類似スコアを計算し，それらの密度分布を利用して閾値を定める．

2. 提案手法

Twitter では，類似スコアが高いツイートの数は類似スコアの低いツイートよりも数が少ない．そのため，ニュースが配信されるより過去に投稿されたツイートとの類似スコアを計算し，ほとんどの類似スコアを超えるような閾値を定めることを考える．それにより，他の多くのツイートが取り得る類似スコアよりも相対的に高い類似スコアを持つツイートのみを取得できる．ニュースが配信された時刻より後に投稿されるツイートは，閾値より高い類似スコアを持つツイートのみを取得することで，ニュースと関連のある可能性が高いツイートのみを集めることができる．

2.1. 類似スコア

本研究では．ニュースとツイートの類似スコアを計算するために [2] で用いられている類似スコアを一部変更した以下の式を用いる．

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

†筑波大学システム情報工学研究科, Graduate School of Systems and Information Engineering, University of Tsukuba
‡筑波大学システム情報系情報工学域, Faculty of Engineering, Information and Systems, University of Tsukuba

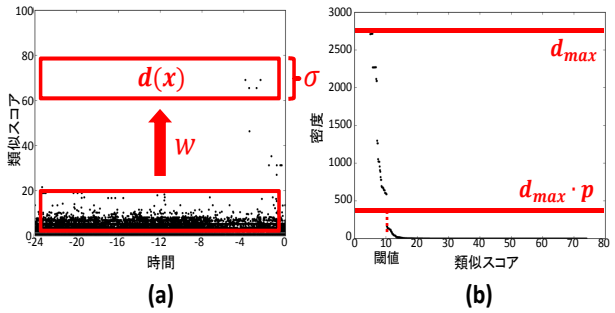

図 1: 閾値の導出

s はニュースを，u はツイートを表している．s_i はニュース内における単語 i の出現回数，$I(u_i)$ はツイート内において単語 i が出現した場合 1，出現しなかった場合は 0 の値をとる関数である．$|s|$ はニュース内の単語数を表している．また，ニュース集合を S とした場合，逆文書頻度 $idf(i)$ は以下の式を用いて計算する．

$$idf(i) = 1 + \log\left(|S|/(1 + |\{s \in S | s_i > 0\}|)\right) \qquad (2)$$

$\{s \in S | s_i > 0\}$ はニュース集合 S の内，単語 i を含むニュースの数を表している．

2.2. 閾値の導出

ニュースが配信された時刻より過去のツイートを用いてニュースに適切な閾値を定めることを考える．図 1(a) は，あるニュースに対する過去 24 時間のツイートを表したものである．各点がツイートを表しており，縦軸がニュースとツイートの類似スコア，横軸がニュースが配信された時間を 0 とした時の相対的な時間を表している．

ニュースが配信された時刻より過去のツイートの内，$score(s,u) > 0$ のツイート集合を U とする．これらのツイート U に対して，類似スコアを x ，ウィンドウ幅を δ とした時，$x - \delta/2$ 以上 $x + \delta/2$ 未満に存在するツイートの数から求まる密度を $d(x)$ と定義する．この時，類似スコア $x = 0$ から最大の類似スコアをもつツイートまでスライド幅 w で密度 $d(x)$ の計算を行う．図 1(a) のニュースに対して，密度 $d(x)$ の計算を行った結果を図 1(b) に示す．この中で最大の密度を d_{max} とした時，類似スコアを昇順に調べ，以下の条件を満たす最初の x を閾値とする．

$$d(x) < d_{max} * p \qquad (3)$$

p は d_{max} からどの程度密度が減少した位置を閾値にするかを決めるためのパラメータである．

2.3. 閾値以上のツイートの取得

ニュースが配信された時刻より後に投稿されたツイートに対しては，ニュースとツイートとの類似スコアを計算し，2.2 節で定めた閾値を上回っているツイートの

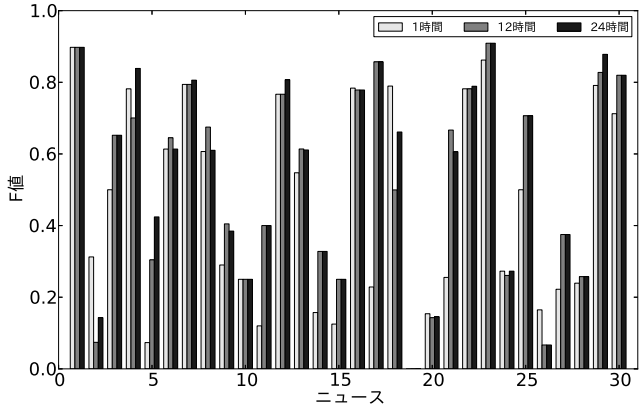

図 2: ツイートの量と F 値

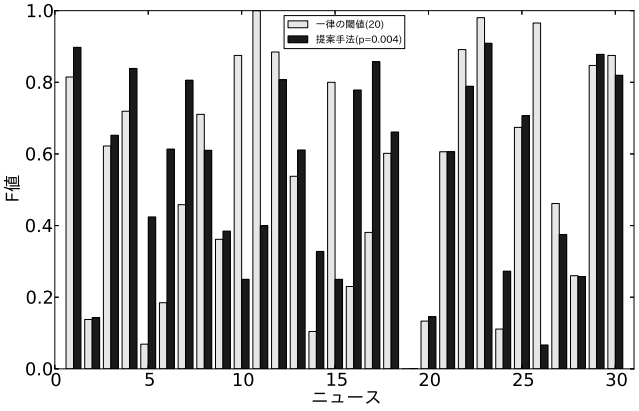

図 3: F 値による手法の比較

みを取得する．計算量を削減するために，[2] で用いられている DAAT for pub-sub with skipping の一部を変更した閾値ベースの DAAT 法を用いるが詳細は省略する．

3. 評価実験

3.1. データセット

実験には 6 月 7 日の 0 時 0 分から 6 月 11 日の 0 時 0 分までに投稿された日本語のツイート 347 万件，6 月 9 日に Yahoo!ニュースの速報ニュースで配信されたニュース 1037 件を用いる．また，IDF の計算には 5 月 1 日の 0 時 0 分から 6 月 1 日の 0 時 0 分までに Yahoo! ニュースの速報ニュースで配信されたニュース 3 万 2 千件を用いる．

3.2. 評価実験

全てのニュースに対して一律の閾値を用いて集めたツイートと，提案手法によって集めたツイートにおいて，どの程度関連のあるツイートを取得でき，どの程度関連のあるツイートを取り損ねているのかを評価する．実験には，ニュース 1037 件からランダムに選択した 30 件のニュースを用いる．それら 30 件のニュースに対し，それぞれのニュースが配信されてから 24 時間以内のツイートの内，類似スコアの高い 1000 のツイートを評価者に評価してもらう．評価は「関連がある」「やや関連がある」「ほとんど関連がない」「関連がない」の 4 段階で，ニュース 1 件に対し 3 人で評価を行う．そのうち,「関連がある」「やや関連がある」の評価が過半数を超えたものをニュースと関連のあるツイートとする．この関連のあるツイートを正解のツイートとして，各手法で取得したツイート集合に対して適合率・再現率・F 値を計算し，結果の比較を行う．また，ニュースが配信されるより過去何時間分のツイートを用いて閾値を決定するかで結果がどのように変化するのかを調査する．

3.3. 結果・考察

図 2 は，ニュースが配信されるよりも前の 1 時間・12 時間・24 時間分のツイートを用いて閾値を定めてツイートを取得した時の結果である．縦軸が集めたツイートに対する F 値，横軸がニュースの ID を表している．12 時間分のツイートを用いて閾値を定めた場合と 24 時間分のツイートを用いて閾値を定めた場合には大きな差は見られなかった．しかし，1 時間分のツイートを用いて閾値を定めた場合は他の 2 つよりも精度が悪く，F 値が低いニュースが 30 件中 16 件存在した．

図 3 はニュース全てに一律の閾値を定めた場合と提案手法を比較したものである．一律の閾値は類似スコアの閾値を 20，提案手法は $p = 0.004, \delta = 10, w = 0.1$ とし，24 時間分のツイートを用いて閾値を定めた．提案手法では 30 件中 17 件のニュースにおいて一律の閾値を定めた場合よりも高い F 値を示した．一律の閾値を定めた場合においては，30 件中 10 件のニュースにおいて提案手法より高い F 値を示した．これは，密度の変化が明確でないニュースの場合，提案手法では関連のないツイートを余分にとってしまうことがあるためだと考えられる．

4. まとめ

本研究では，ニュースが配信されるよりも過去のツイートを用いて閾値を定め，ニュース掲載以降の関連のあるツイートを集める手法を提案した．その結果，一律の閾値を用いるよりも，提案手法のほうが関連のあるツイートを適切に集めることができるニュースが多いことがわかった．今後の展開としては，全く同じ内容のツイートやリツイートが密度の変化に影響を与えているため，これらを考慮した手法の改善を行う．

謝辞

本研究の一部は，文部科学省・未来社会実現のための ICT 基盤技術のの研究開発「実社会ビック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．また，本研究に際して，実験の評価をして下さった研究室の皆様に深く感謝致します．

参考文献

[1] 大西 誠, 北川 博之, 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, 情報処理学会第 76 回全国大会, IPSJ 全国大会 (2014).

[2] 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).