マイクロブログ上の中心的話題とそれに対するユーザの反応の抽出

藤川 智英 † 鍛治 伸裕 ‡ 吉永 直樹 ‡ 喜連川 優 ‡
† 東京大学大学院 情報理工学系研究科 ‡ 東京大学 生産技術研究所

1 はじめに

Twitter などのマイクロブログにおいては日々大量の情報が流れている．その中で具体的何が話題になっており，その話題に関連して人々がどのような発言を行っているのかを知ることは，社会分析をはじめとする各種応用において重要なことである．

本稿では，大規模なマイクロブログ記事の集合における中心的話題を抽出する手法を提案する．具体的には，何らかの話題を形成する語（話題語と呼ぶ）を自動抽出し，互いに関連する話題語のクラスタリングを行うことによって話題抽出を実現する．

また，上記処理によって得られた話題語クラスタに対して，そのクラスタが形成する話題について言及しているマイクロブログ記事を自動的に収集する手法の提案も行う．

2 関連研究

話題解析の研究に最初に取り組んだものとしては，確率オートマトンを用いた [1] が存在するが，この手法は計算的なコストが高く，また,「どれぐらい話題になっているか」を定量的に図ることができない．そこで，[2] では，確率分布を仮定し，確率の低い事象が発生するほど話題になっているとみなす手法が提案されている．

3 話題語の抽出とクラスタリング

3.1 Burst score にもとづく話題語候補の抽出

直近の一定期間（例：30 分）収集したツイートにおいて，ある単語が burst（出現率が急激に上昇）しているかどうかを判断するため，その確率がどれほど低いかを計算することを考える．[2] では単語の出現の確率分布として，二項分布を仮定している．つまり，以下のように p をその単語の平均出現率とすると，

$$p = \frac{\text{その単語が出現したツイートの数}}{\text{全ツイート数}} \tag{1}$$

n 個のツイートを収集した時，そのうち k 個のツイートに単語が出現する確率は，以下の式で計算できる．

$$prob(n,k) = p^k(1-p)^{n-k}\binom{n}{k} \tag{2}$$

ここでは，ある期間収集したツイートにおいて，出現率が平均出現率 p を上回り，かつ，この式で計算した確率が 0.05 以下のとき，その単語が burst している可能性があるとみなす．そして，その burst の度合いを，情報量を用いて以下の式で表す．

$$Score = -\log prob(n,k) \tag{3}$$

3.2 PageRank を用いた話題語の選出

前項では確率が 5 %以下の事象を「burst の可能性がある」とみなしたが，例えば 1 万の単語があれば，そのう 5 %，500 個はたとえ burst してなくても，この 5 %の領域に入ってしまう．そのため，実際に話題になっているものと，偶然 burst しているものを区別する必要がある．何かが話題になる場合，複数の単語が同時に burst し，共起が起こることが多い．

- 例：サッカーの日韓戦
  - 「サッカー」「日本」「韓国」「本田」(選手名)
- 例：「スマトラ沖で地震」というニュース
  - 「スマトラ」「地震」「津波」「マグニチュード」「インドネシア」

そこで，共起している単語があるほどその Score を高くみなすことで，本当に burst しているかどうかを区別することができる．本研究では，これを実現するために，PageRank を応用する．PageRank では，以下の仮定が成り立つとして Web ページの重要性を判断する．

- 多くのページからリンクされているページほど重要なページである．

Extraction of topics and users' attitudes from microblogs
FUJIKAWA TOMOHIDE†, KAJI Nobuhiro, YOSHINAGA Naoki, and KITSUREGAWA Masaru†
† Department of Information Science and Technology, University of Tokyo
† Institute of Industrial Science, University of Tokyo

- 重要なページにリンクされているページほど重要なページである．

単語の話題の度合いの場合も，以下の仮定が成り立つと想定できる．

- 多くの単語と共起している単語ほど話題になっている単語である．
- 話題になっている単語と共起している単語ほど，話題になっている単語である．

ここから，リンクを共起度に置き換えれば，共起度が高いものほど burst score が大きくなるように再計算することができる．実際，PageRank を本来の用途とは異なり，類似度が高いものほどランクを高くするために使用した例として，VisualRank [3] がある．

これにより，他の burst している単語と関係がある本当に burst している単語のランクをあげられる他，多くの他の重要な単語と共起している話題の中心となっている単語のランクを挙げることができる．

このようにして得られた単語のランキング上位 100 語を話題語として抽出する．

3.3 話題語のクラスタリング

クラスタリングは階層的クラスタリングを用いる．ただし，類似度の計算は以下のように行う．クラスタ c, c' が存在した時，$t_i \in C, t'_j \in C'$ とすると，

$$r = \sum_i \sum_j w_i w_j cooc_{ij} \tag{4}$$

ここで，$cooc_{ij}$ は共起頻度，w_i, w_j は 式 3 のスコアをそのクラスタ内での正規化したものであり，ランクの高い単語との共起ほど重要とみなすようになっている．

4 話題に関連するツイートの収集

次に，クラスタと関連のあるツイートを抽出する．クラスタ内の単語を一つでも含むツイートに対して，クラスタ内の単語 t_i の正規化された重み w_i の合計をそのツイートのスコアとし，それが一定の閾値以上ならば関連するツイートであるとみなす．これにより，URL などと関係なく,「中心的な単語が多く含まれているかどうか」のみで話題と関連があるかどうかを判定するため，テレビ番組や一つのツイートが発信源の話題に対しても適用することができる．

5 予備実験: 関連ツイートの判定

実験には，Streaming API で収集したツイートを用いた．A「少女時代，KARA が紅白に内定」B「沖縄防衛局長が女性誹謗発言」C「アメリカン航空が破産法申請」という 3 つのニュースに対して関連ツイートの判定実験を行った．閾値を 0.1 刻みに動かして実験を行ったところ，どの場合も 0.4〜0.6 範囲で F 値が最大となった．閾値を 0.5 とした場合の Precision, Recall, F 値は表 1 の通りである．

表 1: 予備実験結果

	適合率	再現率	F_1 値
A	0.69	0.84	0.76
B	0.84	0.76	0.80
C	0.96	0.80	0.88

6 まとめ

本研究では PageRank の応用などを用いて burst している単語のクラスタを抽出し，それと関連するツイートを収集する手法を提案した．将来的な課題としては，まず，共通の単語が多く含まれる似たような話題に対してはクラスタリングが上手くいかないことがあげられる．また，関連ツイートの収集が，ニュース以外の話題についてもうまくいくかどうかなどの点について研究を進めたい．

参考文献

[1] Jon Kleinberg. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, Vol. 7, No. 4, pp. 373–397, October 2003.

[2] Gabriel Pui Cheong Fung, Jeffrey Xu Yu, Philip S. Yu, and Hongjun L. Parameter free bursty events detection in text streams, 2005.

[3] Yushi Jing and Shumeet Baluja. Visualrank: Applying pagerank to large-scale image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 11, No. 30, pp. 1877–1890, November 2008.