4N-4

Twitterの時系列解析による注目話題の抽出

Extracting Featured Topics by Chronological Analysis of Twitter

木原 大志† 白木原 渉†† 大石哲也‡ 越村 三幸‡‡ 藤田 博‡‡ 長谷川隆三‡‡

Hiroshi Kihara Wataru Shirakihara Tetsuya Oishi Miyuki Koshimura Hiroshi Fujita Ryuzo Hasegawa

†九州大学工学部電気情報工学科 ††九州大学大学院システム情報科学府

‡九州大学情報基盤研究開発センター ‡‡九州大学大学院システム情報科学研究院

1 はじめに

ユビキタスの到来により，多くの人が自由に情報を発信し，それを自由に入手できるようになった．特に情報発信においては，スマートフォンなどの高機能な携帯電子端末の台頭や，SNSやブログといった様々なWebサービスの普及により，より気軽に，かつ即時性を持って行われるようになってきている．

しかし，この傾向は同時に情報検索を困難にもしている．情報発信者が増えたことで単純な情報量が増え，対応するためには相応の計算資源が必要である．また個々の発信が簡便になることで，それぞれの情報の価値に差が生まれ，その中から有意義な情報のみを得ることが難しい．

本論文では、このような日々更新される膨大な情報群から，有意義な情報の例として，世間での流行や，注目されている話題を抽出することを目的とする．

2006年に開始し，近年日本でも普及しつつあるWebサービスTwitter[1]は， 情報発信，情報交換が非常に容易であり，上記のような「気軽さ」「即時性」に優れたサービスである．このTwitterから注目されている話題を抽出するために，ツイートに付随する時間情報に着目して[1]，単語ごとの特徴抽出を行う．

2 時系列データの評価方法

ある時間 $t_i \leq t \leq t_{i+1}$ の間に採取されたツイートの数を u_i，これらのツイートの内，単語 w を含むツイートの数を x_i とする．$1 \leq i \leq n$ の間ツイートを採取すれば，n次元のベクトル $\vec{x}$，$\vec{u}$ が求まる．そこで $\vec{x}$ と $\vec{u}$ との関係性について議論する．

もし $\vec{x}$ と $\vec{u}$ との相関が強ければ,「常にある一定の割合で出現する単語」と位置付けることができ，すなわち抽出するに値しない単語であると言える．逆に相関が弱いのであれば，通常とは異なる時系列をたどったということであり，何かしらのニュース性をはらんだ単語である可能性が高い．

$\vec{x}$ と $\vec{u}$ との関係性を測定するにあたって，以下の指標を用いる．

[1]Twitter https://twitter.com/

一般的なものとして，ピアソンの相関係数（式1），コサイン類似度（式2）がある．これらは「類似度」であり，値が大きいほど二つの系列は類似していて，値が小さいほど類似していないことを表す．

$$corr(\vec{x},\vec{u}) = \frac{\sum (x_i-\bar{x})(u_i-\bar{u})}{\sqrt{\sum (x_i-\bar{x})^2}\sqrt{\sum (u_i-\bar{u})^2}} \quad (1)$$

$$cos(\vec{x},\vec{u}) = \frac{\vec{x}\cdot\vec{u}}{|\vec{x}|*|\vec{u}|} = \frac{\sum x_i u_i}{\sqrt{\sum (x_i)^2}\sqrt{\sum (u_i)^2}} \quad (2)$$

これらに加え，以下の式による指標3（式3），指標4（式4）を導入する．これらの指標は「距離」であり，値が大きいほど二つの系列の間の関連が薄いことになる．

$$PV(p) = \frac{1}{n}\sum (p_i-\bar{p})^2 \quad (3)$$

$$(\text{ただし}\vec{p} = (p_1, p_2, \cdots, p_n), p_i = x_i/u_i)$$

$$CHPM(x,u) = \frac{max(|x_i-u_i|)}{\sum |x_i-u_i|} \quad (1 \leq i \leq n) \quad (4)$$

指標3は，単語の出現数の時系列 $\vec{x}$ を，全体の時系列 $\vec{u}$ で正規化し，その標本分散をとっている．正規化することで,「全体に対する比率」の系列となり，その散らばり具合を計算することで，全体の流れに逆らった系列を持つ単語が導出できる．

指標4は，チェビシェフ距離とマンハッタン距離の比を用いている．各時間における単語の出現数とツイート総数の差について，マンハッタン距離はその総和であり，対してチェビシェフ距離は，その中の最大値である．これらの比をとることで，ツイート数が大きいタイミングで多くない単語，裏を返せばツイート数が多くないタイミングで突出する単語を抽出できる．

3 実験と結果

データセットとして，2011年12月19日午前10時30分から，翌20日午後10時30分までの24時間で言語を日本語に設定しているユーザによるツイートを採取した．この日に起こった大きな事件として，北朝鮮の金正日総書記の死去[2]が挙げられる．

[2]http://mainichi.jp/select/world/graph/20111219/

表 1: 相関係数，コサイン類似度による話題性

	相関係数		コサイン
遅刻	-0.2631	ikamusume	0.32486
今朝	-0.25571	ジョンイル	0.47013
寝坊	-0.2405	テレ	0.51521
6	-0.22372	遅刻	0.51787
アダルト	-0.18345	今朝	0.52072
早起き	-0.1793	キム	0.5305
天気	-0.13124	死亡	0.5411
気温	-0.09622	寝坊	0.55296
ジョンイル	-0.04901	お昼	0.5751
お昼	-0.02922	正日	0.57648
キム	-0.0153	死去	0.58853
死亡	-0.00356	早起き	0.60128
弁当	0.00737	イカ	0.60695
テレ	0.02458	書記	0.61764
死去	0.02544	朝鮮	0.61926

表 2: による話題性

	PV		CHPM
正日	1.44E-04	正日	0.16885
書記	1.70E-05	書記	0.1681
死去	1.63E-05	死去	0.16809
12	1.32E-05	12	0.16807
死亡	1.30E-05	北朝鮮	0.16805
北朝鮮	1.13E-05	学校	0.16805
20	8.23E-06	定期	0.16803
学校	6.14E-06	死亡	0.168
2011	5.88E-06	仕事	0.168
ジョンイル	5.75E-06	2011	0.16798
電車	5.19E-06	19	0.16794
キム	5.00E-06	授業	0.16792
仕事	4.81E-06	電車	0.16792
風呂	4.40E-06	ジョンイル	0.16792
クリスマス	4.13E-06	キム	0.16792

採取されたツイートを形態素解析エンジン MeCab[3]に入力し，名詞として判定された文字列をカウント，上述の４手法で時系列の特徴を抽出した．

表１および２は，各指標によってより特徴的であると判断される単語の上位１５件ずつを表示している．(ただし，出現数が１２０件以上のものに限った.)

相関係数の上位３件は，ツイートの数がピークになる午後８時から深夜にかけては出現せず，翌朝５時ごろから出現数が増え始めた単語である．想定した「全体と異なる時系列」ではあるが，その根拠が単に「朝が来た」というだけのことであり，本研究の目的である「ニュース性を持つ単語」ではない．

コサイン類似度では，深夜の TV 番組を象徴する単語（ikamusume，イカ)[4]が増えている．１日の出現数は，全 292,693 件の内わずか 150 件であるが，放送時間付近に集中して出現していた．

指標３および４では，上位３件が共通してこの日のニュースを象徴する単語になっている．上位１５件で比較すると，ニュースに関連する単語が多く抽出できていると言える.「正日」という単語はこの１日で 4795 件に出現し，純粋な出現数が最も多い．指標３の「クリスマス」(2124 件) や，指標４の「定期」(1203 件) など，単純に出現数が多い単語はより抽出されやすいと言える．ただし，やはり単純な出現数の多い単語「お願い」(2274 件)「大丈夫」(2258 件) などは下位にあるため,「出現数は多いが特徴的でない単語」を除去している，という評価もできる．

[3]http://mecab.sourceforge.net/
[4]http://www.ika-musume.com/

4 おわりに

今回は，時系列による特徴抽出という観点から，世間で注目されている話題を抽出するための手法を提案した．各指標によって抽出される単語の傾向が異なるという結果が得られた．

今後はより長期間の標本を用いたり，複数の指標を組み合わせるなどの方法で精度を上げてゆく．また，既存のトレンド抽出手法との比較や，人手による評価など，精度の評価についても研究する予定である．

謝辞 本研究は科研費 (21500102) の助成を受けたものである．

参考文献

[1] 和泉諒, 西山裕之マイクロブログの時系列情報を利用した関連語発見手法に関する研究全国大会講演論文集 2011(1), 681-683, 2011-03-02

[2] 白木原渉, 大石哲也, 長谷川隆三, 藤田博, 越村三幸 Twitterにおける流行語先取り発言者の検出システムの開発 Trendspotter Detection System for Twitter 情報処理学会研究報告. データベース・システム研究会報告 2010-DBS-150(2), 1-8, 2010-07-28

[3] 阿部秀尚，津本周作時系列テキストマイニングによる類似用語の語彙体系内距離の比較
A Comarison of Similarity of Technical Terms in Temporal Patterns on MeSH The 25th Annual Conference of the Japanese Society for Artificial Intelligence, 2011 1D1-3