3S-4

ユーザ間のTweet頻度偏り補正に基づくローカルバースト検出手法

東口 大樹 † 仲野 雅幸 † 佐野 博之 † 白松 俊 † 大囿 忠親 † 新谷 虎松 †
名古屋工業大学大学院工学研究科情報工学専攻 †

1 はじめに

近年，行政のオープン化を目指してある議題に対してインターネットを通じて市民の意見を求め，それを政治に反映しようという試みが行われている．このような試みの一つとして，アイディアボックスがある．アイディアボックスとは経済産業省が運営するウェブサイト「オープンガバメントラボ」のサービスで，参加者は募集テーマに対して，新規アイディアの投稿や，既に投稿されたアイディアへのコメント，賛否の投票などを行うことができる．このように積極的に市民に意見を求め，行政への市民参加を促すことで行政のオープン化を目指している．このような行政のオープン化の動きは地方自治体においても始まっており，今後このような行政のオープン化が更に進むと思われる．このようなアイディアボックスを始めとする意見投稿サイトにおいて，議論の話題はあらかじめ人手で準備しておかなければならず，運営を行う上で負担となるという問題点がある．

本稿では，地方自治体が意見投稿サイトを運営する場合において議論のきっかけとなるように地域トレンドを自動で収集，提示することを考える．地域トレンドはシステムを利用する地方自治体に関係したものでなければならない．そこで本稿では，現在多くの人が利用しているマイクロブログサービスである Twitter を用いて地方自治体に関する情報を集め，その情報の中から地域限定でバーストしている情報（ローカルバースト）を取り出すことによりその地域で話題になっている単語を提示することを考える．Twitter は情報のリアルタイム性に優れ，最新の話題に対していち早く情報を集めることが可能であることや，情報伝播力が大きいこと，更にいつでも，どこでも，誰でも，情報の発信・収集が可能であるという簡易さから有益な情報がいち早く得られると考える．

本研究では，Tweet と地域名の紐付けを行うために本研究室において構築された補集合ナイーブベイズ分類器を用いる．学習には Wikipedia の記事を用いる．Wikipedia 記事中には，Infobox と呼ばれる基礎情報に地域情報（施設の所在地等）が明示されている場合がある．また，Infobox が無い記事の場合でも，記事の 1 文目に地名が現れている場合はその地名と関連がある記事であると考えられる．本研究ではこのような条件を満たす 145,668 記事を抽出し，学習に用いた．この分類器はドキュメントを入力として与えた時，47 都道府県，1788 市町村，190 区の自治体名に分類することができるものである．

Tweet t を 分類する際の地域 C のスコアは，下記のようになる．ただし，w は単語である．

$$score(t,C) = logp(C) - \sum_{w \in t} TF(w,t) logp(w|\hat{C}) \quad (1)$$

145,668 記事から補集合ナイーブベイズモデルを学習するために，スケーラブルに動作する分散環境 Hadoop 上のマイニングライブラリ Mahout を用いた．

Local Burst Detection Based on Weighting Users According to Difference of Tweet Frequency

Daiki HIGASHIGUCHI, Masayuki NAKANO, Hiroyuki SANO, Shun SHIRAMATSU, Tadachika OZONO, and Toramatsu SHINTANI
Dept. of Computer Science, Nagoya Institute of Technology

2 関連研究

本研究においてバーストとは連続して文書が出現するドキュメントストリームにおいてある単語 w が急に文書内に多く出現する様子を表す．ドキュメントストリーム内におけるバースト検出手法として Kleinberg のバースト検出手法 [1] がある．この手法は文書が完全にランダムに到着すると仮定したドキュメントストリームにおいてバーストの検出を行う手法である．また，藤木ら [2] は Kleinberg の手法を，ブログなどの文書数が急上昇しているものや，電子掲示板など一日でも時間帯によって文書数が異なるものに応用するための拡張手法を提案している．これらの手法を用いることでブログや掲示板などのコンテンツにおいてバーストの検出を行うことができる．

Kleinberg の手法 [1] において λ は単位時間あたりの文書数，つまりある時間区間を T として，その時間区間内に K 個の文書が出現する時，次の式で表すことができる．

$$\lambda = K/T \quad (2)$$

Kleinberg の手法において文書の出現間隔 x がランダムであると仮定をしていることは x が指数分布に従うということと等価である．よって x の確率密度関数は以下のようになる．

$$f(x) = \lambda e^{-\lambda x} \ \ (\lambda > 0) \quad (3)$$

Kleinberg はバーストをある程度の固まりとして抽出するために，平常状態 0 とバースト状態 1 の 2 状態オートマトンを考えて，それぞれの状態の状態遷移コストを計算することにより，バースト状態の検出を行なっている．この時，平常状態の確率密度関数を $f_0(x) = \lambda_0 e^{-\lambda_0 x}$，バースト状態の確率密度関数を $f_1(x) = \lambda_1 e^{-\lambda_1 x}$ とすると，パラメータ s を用いて $\lambda_1 = s\lambda_0$ とすることによりバースト検出の度合を調整する．本稿では藤木ら [2] が用いた以下の式を用いてコストを計算する．ただし，t はドキュメントストリームにおけるあるドキュメントを表し，j は 2 状態オートマトンである平常状態 0 かバースト状態 1 のいずれかであり，l は一つ前のドキュメント $t-1$ における 2 状態オートマトンの値を表し，τ は $l=0$，$j=1$ の時にのみ γ となる状態遷移コストであり，バースト状態 1 に容易にならないようにするためのものである．

$$C_j(t) = -ln f_j(x_i) + \min_l (C_l(t-1) + \tau(l,j)) \quad (4)$$

ドキュメント到着時に上式を用いて $C_0(t)$ 及び $C_1(t)$ を計算し，$C_1(t)$ の方が小さくなる時にバースト状態であるとし，その時の $C_0(t) - C_1(t)$ をバースト値，複数の Tweet から構成されるバースト部分に対しては，その部分のバースト値を全て足しあわせたものをバースト値スコアとして，これらのバースト値，及びバースト値スコアを用いてランク付けを行っている．

藤木らは Kleinberg の手法を電子掲示板に対して用いる上で，時間帯によって平常状態のドキュメント数が異なることを挙げ拡張手法を提案している．ある電子掲示板に対して一日に n 個の書き込みが観測された場合を考え，一日を N 分割したときのそれぞれの時間区間 i での平均書き込み数を K_i,

時間区間 i の長さを T_i とする時，各時間区間 i に対する λ_0 を次式のように計算している．

$$\lambda_{0,i} = \frac{n \times (K_i / \sum K_i)}{T_i} \tag{5}$$

3 ローカルバーストの検出

本稿ではローカルバーストの検出を行う際に生じる問題として，地域によって生じる Tweet 数の違いと，ユーザ数が少ない地域における一人辺りの Tweet 数の違いを考慮する．

3.1 Tweet 数に応じた計測時間間隔の補正

バーストの検出を行う際，一定以上の Tweet 数が必要である．しかし，地域によって Tweet 数が異なることから，各地域によってどのぐらい Tweet が行われているのかを考慮しなければならない．バーストの検出を行う場合に最低限必要となる Tweet の数が K_0 である時に，地域 c において観測時間 T' 分あたり Tweet が K_c 個到着したと考えると，$\lambda = \frac{K_c}{T'}$ なので，次の式のように計測時間間隔 T_c を補正すればよい．

$$T_c = \frac{K_0}{\lambda} \tag{6}$$

$$= \frac{K_0 T'}{K_c} \tag{7}$$

これにより，Tweet 数が少ない地域 c では時間間隔 T_c が長く補正される．

3.2 ユーザ数に応じた発言の重みの補正

Twitter ユーザは多く存在するが，ユーザの多くは都市圏に偏っている．そのため，Twitter のユーザ数が極端に少ない地域も存在する．これらの地域ごとのユーザ数の違いから Twitter のユーザ数の少ない地域の場合，多くの Tweet を行うユーザがその地域全体の Tweet に占める割合が多くなることが考えられる．この場合，そのユーザの地域のバースト検出時に，そのユーザの発言が多く含まれることによりバースト検出に与える影響が大きくなり，Tweet 数の少ない他のユーザの発言がバースト検出の結果に反映されにくくなってしまう．そこでユーザ数の少ない地域においては，ドキュメントストリーム内における Tweet 数の多いユーザの重み付けを小さくすることによりこれらの問題を解決する．

ある地域 c においてある期間内で単語 w を含む Tweet を行っているユーザ数を $u_{c,w}$ として，あるユーザの単語 w を含む Tweet 数を x とする．この時，単語 w のバースト計算時のそのユーザの発言の重み $weight(u_{c,w}, x)$ をパラメータ β を用いて次式のように決める．

$$weight(u_{c,w}, x) = x^{-\frac{\beta}{u_{c,w}}} \tag{8}$$

この重み $weight(u_{c,w}, x)$ は図1 で表されるように，ユーザ数が多い地域の場合は Tweet 数が多くなってもそれほど重み付けが変わらないが，ユーザ数が少ない地域の場合は Tweet 数が多くなるにつれて重み付けが小さくなるようになっている．

単語 w を含む Tweet において実際にバースト値を求める場合，w を含む Tweet をしたユーザの集合を i_w，w を含む Tweet の数を K_w とすると単語 w のバースト値にかけられる重み $weight_w$ は次式のようになる．

$$weight_w = \frac{\sum_{i_w} x_{i_w} weight(u_{c,w}, x_{i_w})}{K_w} \tag{9}$$

この $weight_w$ をバースト値スコアに掛けることにより，ユーザ数に応じた発言の重みの補正を行う．

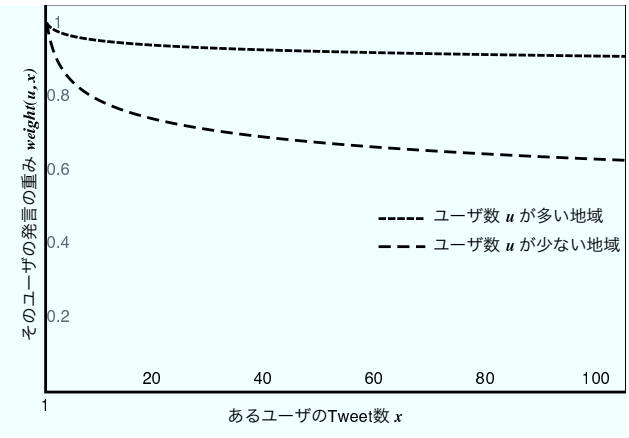

図 1: ユーザの地域と Tweet 数によるバーストの重み付け

4 考察

本研究における手法におけるパラメータである K_0 と β については $K_0 = 10, 20, 50$，$\beta = 1, 10, 100$ を用いて予備実験を行うことで，最適な出力が得られるようにする．

Kleinberg の手法には２つのパラメータがあり，Kleinberg は $s = 2$，$\lambda = 1$ としている．それに対して藤木らはトピックワードリストを求める際の予備実験で $s = 4$，$\lambda = 2$ で良い結果が得られたとしている．この場合，s を大きくすることによってバースト判定を厳しくする効果がある．本研究においては藤木らの決めたパラメータセッティングを踏襲し，$s = 4$，$\lambda = 2$ を用いる．

本研究における手法を単語だけでなく，Tweet 内に含まれる URL に適用することにより URL のローカルバーストの検出も行うことができる．URL のローカルバーストを検出できると，ある地域において注目の集まっている Web ページを抽出することができ，この Web ページを提示することができることで単語だけでなく Web ページも意見収集に用いることができる．

本システムの応用例として，本システムの出力をスレート型端末に読み込んでバーストする単語や記事を端末のユーザに提示することにより，単語や記事に関連する住民の意見を入力するインターフェースが作成できると期待される．

5 おわりに

本論文では，ローカルバーストを検出する手法として Tweet 数に応じた計測時間間隔の補正手法とユーザ数に応じた発言の重みの補正手法について提案した．本手法を用いてローカルバーストの検出を行うことで各地域において話題となっている単語を抽出することができる．また，同様の手法で Tweet 内に含まれる URL においてもローカルバーストの検出が行える．これらのローカルバースト検出手法を用いて，地域におけるトピックワードのランク付けを行い，これを提示することにより地域における議論への話題提供ができると期待される．

参考文献

[1] Jon Kleinberg: ”Bursty and hierachical structure in streams”, Proc. the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002.

[2] 藤木稔明, 南野朋之, 鈴木泰裕, 奥村学: document stream における burst の発見, IPSJ SIG Notes 2004(23), 85-92, 2004.