クラスタへの帰属度を考慮した位置に基づくバースト検出手法の検討
A Study on Location-based Burst Detection Considering Membership

事崎　翔太
Shouta Kotozaki
広島市立大学
情報科学部
Email: t20076@edu.ipc.hiroshima-cu.ac.jp

田村　慶一
Keiichi Tamura
広島市立大学大学院
情報科学研究科
Email: ktamura@hiroshima-cu.ac.jp

北上　始
Hajime Kitakami
広島市立大学大学院
情報科学研究科
Email: kitakami@hiroshima-cu.ac.jp

Abstract—**Nowadays, a large number of geo-referenced documents, i.e., text messages including location information, are posted on social media sites. People transmit and collect information over the Internet through these geo-referenced documents, which are usually related to not only personal topics but also local topics and events. Detecting local topics and events in geo-referenced documents is beneficial for many different geo-mobile application domains. Burstiness is one of the simplest and most effective criteria for extracting hot topics and events. In this paper, we propose a novel clustering-based burst detection method that integrates fuzzy c-means clustering method into location-based burst detection. Moreover, to detect important bursts, we consider the membership of documents to clusters. To evaluate the proposed burst detection algorithm, we used an actual set of geo-referenced documents, which are crawling tweets posted on the Twitter site. The experimental results show that our new burst detection algorithm can detect location-based bursts for topics considering membership.**

I.　はじめに

インターネット上で生成される文書データを時系列に到着するストリームデータ（以下，文書ストリームと呼ぶ）として扱い，文書ストリーム上に頻繁に現れてくる事象（ユーザ，語句や場所など）のバーストを検出する手法が盛んに研究されている．ある事象の出現頻度が通常の出現頻度と比較して多く，また，急激に増加している現象をバーストという．文書ストリーム上においてバーストを検出することで，インターネット上でユーザの関心の高い事象を検出することができ， 社会的な話題の分析，観光情報，情報推薦などに応用することができる．

文書ストリーム上において文書データに含まれる事象のバーストを検出する手法として Kleinberg のバースト検出アルゴリズム[1]が提案されている．Kleinberg のバースト検出アルゴリズムでは，あるキーワードを含む文書データの到着間隔に着目し，到着間隔が短くなっている区間をバースト状態であると判定し，バースト状態である区間を検出することができる．ただし，ただ単に到着間隔が短くなっている部分を検出するだけだと細かい区間のバースト状態が点在するようになるため，Kleinberg のバースト検出アルゴリズムではバーストを状態遷移として捉え，まとまりのある区間でバーストを検出することができる．

一方，近年，GPS 付き携帯情報端末やスマートフォンの普及とともに，インターネット上で生成される文書データには，文書データが生成された時間だけではなく，文書データが生成された位置に関する情報（位置情報）が付与されるようになってきている[2]．位置情報が付与された文書データの内容は，様々な時間に様々な位置で人々が目にしたことなど，位置に関連するイベントやトピックと結びついている可能性が高い．よって，位置を考慮してバーストを検出することが重要となる．

我々は先行研究[3]として，位置情報が付与された文書データから構成される文書ストリームにおいて，位置に基づくバーストを検出する手法をしている．例えば，ユーザの近くで発生した事象に対しては，バースト状態を強くし，ユーザの遠くで発生した事象に対してはバースト状態を小さく提示することができる．しかしながら，先行研究では，文書データに含まれるキーワード毎のバーストを検出可能であるが，トピック毎のバーストを検出することができないという課題が存在する．

そこで，本研究では，文書データをファジイ c-means 法[4]により分類し，クラスタ毎にバーストを検出することでこの問題点を解決することを試みる．文書データをファジイ c-means 法により分類することでトピック毎にバーストが検出可能となる．また，クラスタへの帰属度とユーザとの距離を用いて文書データの影響度を算出し，影響度を用いて Kleinberg のバースト検出アルゴリズムで検出されるバーストの補正を行う．例えば，ユーザからの距離が遠い文書データと帰属度が小さい文書データの影響を排除でき，ユーザ周辺のトピックのバーストが検出可能となる．

本論文の構成は以下の通りである．第 2 章では，Kleinberg のバースト検出アルゴリズムについて説明する．第 3 章で提案手法を述べ，第 4 章で評価実験について報告し，第 5 章で本論文のまとめを行う．

II. バースト検出

文書ストリームとは，文書データ d_t が到着した後，x_t の間隔において次の文書データ d_{t+1} が到着するというような時系列の文書データ集合からなるストリームのことを示す（図 1）．社会的に関心の高いトピックが発生すると，そのトピックを表すキーワードを含む文書データは次々生成される．そして，生成される文書データ間の到着間隔は短くなる．到着間隔が短くなっている区間をバースト状態として定義したのが Kleinberg である．

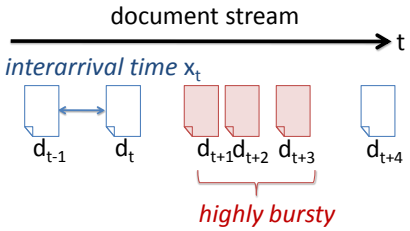


図 1 文書ストリーム

Kleinberg のバースト検出アルゴリズムでは，文書データの到着間隔 x_i は隠れマルコフモデルの内部状態に応じて確率的に出力される記号であるとみなす．そして， m 個の状態を持つ隠れマルコフモデルを仮定する．ここで，バースト検出は文書到着間隔列 $x = (x_1, x_2, ..., x_n)$が与えられた時，次のコスト式を最小にする最適の状態遷移列 $s=(s_1,s_2,..,s_n)$ (各 s_i は状態番号を表す)を求める問題として定義される．

$$C(s|x) = \left(\sum_{i=1}^{n-1} \tau(s_i, s_{i+1})\right) + \left(\sum_{i=1}^{n} -\ln f_{s_i}(x_i)\right). \quad (1)$$

式(1)の第一項は，内部状態が状態 s_i ら s_{i+1} に遷移する際のコストの総和であり，関数 τ は状態 i から状態 j に遷移に必要となるコストを返す関数であり，次のように定義される．

$$\tau(i,j) = \begin{cases} (j-i)\gamma, & if\ j > i, \\ 0, & otherwise. \end{cases} \quad (2)$$

上記の式では，高い状態への状態遷移はユーザが指定したパラメータ γ に比例したコストが必要となり，低い状態への状態遷移のコストは 0 となっている．また，関数$f_k(x_i)$は状態 k で到着間隔 x_i を発生するために必要なコストであり，第二項はその総和となる．

$$f_k(x_i) = \lambda_k e^{-\lambda_k x_i}. \quad (3)$$

ここで， λ_k は以下のように定義される．n は観測時間 T に到着した文書データの数であり，n/T は単位時間当たりの文書データ数を示す．

$$\lambda_k = \frac{n}{T}\beta^k. \quad (4)$$

$C(s|x)$ を最小とする最適の状態遷移列 $s=(s_1,..,s_n)$は，隠れマルコフモデルに対するビタビアルゴリズムを用いて次の動的計画問題を解くことで求めることができる．

$$C_j(i) = -\ln f_j(x_i) + \min_l(C_l(i-1) + \tau(l,j)) \quad (5)$$

式（5）では，系列の i 番目において状態 s_j で終了するときの最小コストを $C_j(i)$で表している．つまり，$C_j(i)$は直前の$(i\text{-}1)$番目の系列における最小コスト $C_l(i\text{-}1)$と $\tau(l,j)$から求めることができる．また，最適な状態遷移列は，最小な$C_j(n-1)$からコストが最小となる経路をトレースバックすることで取り出すことができる．

III. 提案手法

本章では，文書ストリームのデータモデル，問題設定と文書データの影響度を述べ，提案手法のアルゴリズムを説明する．

A. データモデル

文書ストリーム上の文書データ sd_i は 3 つのデータ要素 $sd_i = <text_i, t_i, p_i>$から構成されるものとする．ここで， $text_i$ は当該文書データの内容（タイトルやテキストデータなど），t_i は当該文書データの生成時刻， p_i は位置情報（経度・緯度）である．また，文書ストリーム上で到着した文書データ系列を $SDS = (sd_1, sd_2, ..., sd_n)$と表す．また，生成時刻を文書ストリーム上では到着時刻として考える．

B. 問題設定と文書データの影響度

本研究では，文書ストリーム $SDS = (sd_1, sd_2, ..., sd_n)$において，位置情報に着目し，ユーザ周辺に存在するイベントとトピックについて位置に基づくバーストを検出することを目指している．トピック毎にバーストを検出するために，クラスタリングにより文書データをまとめ，クラスタ毎にバーストを検出する．ただし，各文書データは複数のトピックを含む可能性があるため，ファジイ c-means 法を用いてクラスタリングする．

また，先行研究[3]において，位置に基づくバースト検出手法を提案している．この手法では，文書データの影響度をユーザとの間の距離と定め，影響度によりバーストの度合いを補正する．本研究では，クラスタ毎にバーストを検出するために，文書データの影響度として，クラスタへの帰属度とユーザとの間の距離の 2 つを用いる．

$$drate(d,u) = w_d \times \left(\alpha^{\frac{d-dmin}{dmax-dmin}}\right) + w_u \times \vartheta^{1-u} \quad (6)$$

ただし，d はユーザと文書データ間の距離，u はクラスタへの帰属度，$dmin$ と $dmax$ はユーザと文書データ間の距離の最小値と最大値，α とϑは影響度の逓減係数，w_d と w_u は重み係数で，$w_d + w_u$ =1.0 とする．

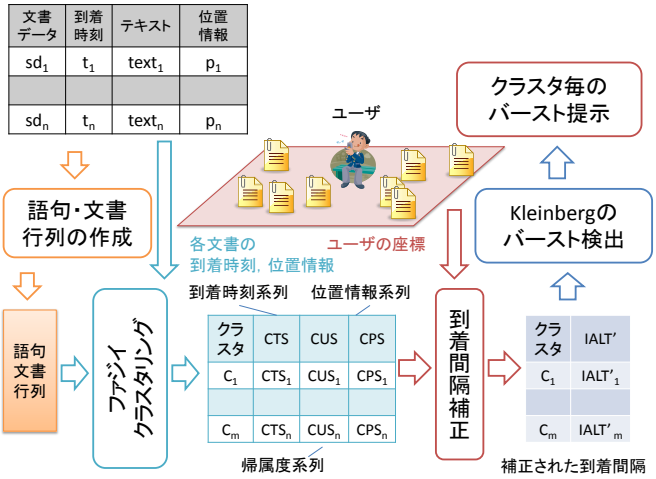


図 2.　提案手法の処理手順

C. アルゴリズム

図 2 に提案手法の処理手順を示す．以下，それぞれの手順の詳細を説明する．

最初に，文書データに含まれる語句を T={$term_1$, $term_2$, ..., $term_m$}とし，語句文書行列 W を作成する．語句文書行列 W の各要素の定義は以下の通りである．α_i は語句の重要度を示す．

$$w_{i,j} = \begin{cases} \alpha_i & \text{if } term_i \text{ is included } text_j \\ 0 \text{ if } term_i \text{ is not included } text_j \end{cases} \quad (7)$$

語句文書行列の各列を文書データの特徴ベクトルとして，ファジイ c-means 法を用いて，文書データを m 個のクラスタ C_i (1≦i≦m)にクラスタリングする．ここで，各文書データ sd_jのクラスタ C_i への帰属度を $u_{i,j}$ とする．ここで，帰属度がユーザの与えた閾値 min_mem より低い文書データを取り除いた文書データをクラスタ i の文書データ系列は，

$$CSDS_i = (sd_{\varphi_i(1)}, sd_{\varphi_i(2)}, \dots, sd_{\varphi_i(numd(i))}) \quad (8)$$

となる．ここで，φ_iは$CSDS_i$を構成する文書データ集合から SDS を構成する文書データ集合への単射であり，$numd(i)$はクラスタ i の文書データ数を返す関数である．また，文書データの到着時刻，クラスタへの帰属度，位置情報を取り出して到着時刻ごとに並べ変えた系列は，

$$CTS_i = \left(t_{\varphi_i(1)}, t_{\varphi_i(2)}, \dots, t_{\varphi_i(numd(i))}\right),$$
$$CUS_i = \left(u_{i,\varphi_i(1)}, u_{i,\varphi_i(2)}, \dots, u_{i,\varphi_i(numd(i))}\right),$$
$$CPS_i = \left(p_{\varphi_i(1)}, p_{\varphi_i(2)}, \dots, p_{\varphi_i(numd(i))}\right), \quad (9)$$

となる．

各クラスタの到着間隔の系列を$IALT_{C_i} = (ialt_{i,1}, ialt_{i,2}, \dots, ialt_{i,|C_i|})$とする．各要素は次の式で求めることができ，文書ストリームの開始時刻を$stime$とする．

$$ialt_{i,j} = \begin{cases} t_{\varphi_i(j)} - stime, & j = 1, \\ t_{\varphi_i(j)} - t_{\varphi_i(j-1)}, & otherwise. \end{cases} \quad (10)$$

ここで，各クラスタの到着間隔を，クラスタへの帰属度とユーザからの距離で補正を行う．Kleinberg のバースト検出アルゴリズムでは，バーストの状態遷移列は到着間隔により決定する．到着間隔を補正することでバーストの度合いを変化させることができる．

$$ialt_{i,j}' = ialt_{i,j} + total\_time \times \left(1 - \text{drate}\left(dist\left(user, p_{\varphi_i(j)}\right), u_{i,\varphi_i(j)}\right)\right) \quad (11)$$

ただし，$total_time$は総時間間隔とする．

各クラスタの各クラスタの到着間隔の系列を$IALT_{C_i}' = (ialt_{i,1}', ialt_{i,2}', \dots, ialt_{i,|C_i|}')$ を入力として，以下のコスト式を最小化する状態遷移系列を求めると，各クラスタのバーストの変化が状態遷移列として抽出される．

$$C\left(s \middle| IALT_{C_i}'\right) = \left(\sum_{i=1}^{n-1} \tau(s_i, s_{i+1})\right) + \left(\sum_{i=1}^{n} -\ln f_{s_i}\left(ialt_{i,j}'\right)\right). \quad (12)$$

IV.　評価実験

評価実験では，2009 年の台風 18 号（図 3 に経路を示す）に対して Twitter でつぶやかれているツイートを twiphoon から収集し，データセットとして用いた．ただし，台風の状況に関係のないものや，宣伝などの記事はあらかじめ除外し，504 件のツイートを用いて実験を行った．ファジイクラスタリングは形態素解析から抜き出した名詞，形容詞，動詞を元にクラスタ数を 50，閾値 min_mem=0.5，w_d=0.5，w_u=0.5，α=0.9，ϑ = 0.3として実験を行った．またユーザの位置として，福岡，大阪，名古屋，東京の主要都市を設定した．

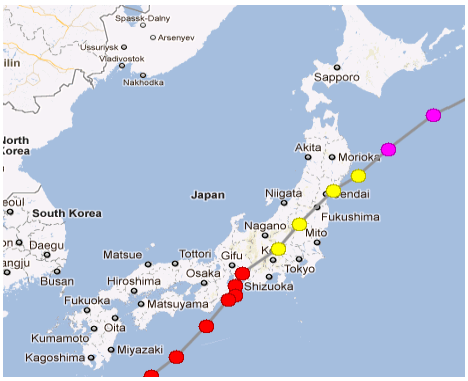
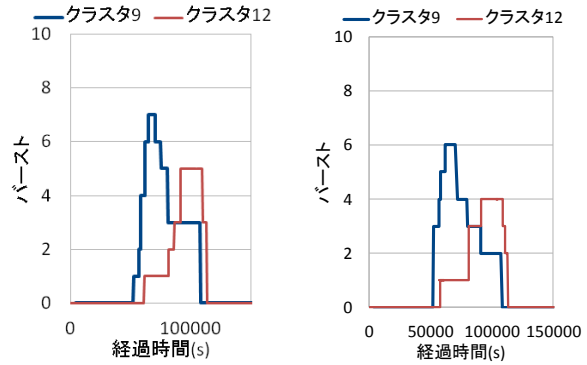
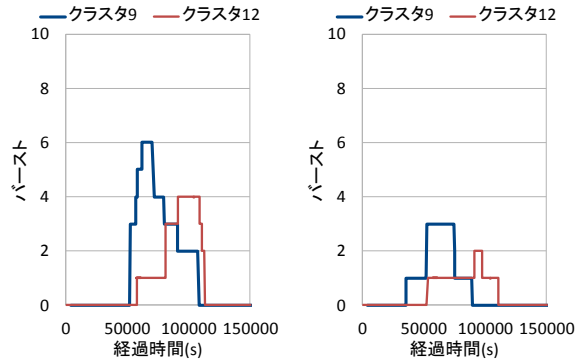


図 3.　2009 年台風 18 号の経路

帰属度の平均値が高い，上位 2 件のクラスタ（クラスタ 9 とクラスタ 12）についてバーストの検証を行った．クラスタ 9 には，186 個のツイートが含まれ，帰属度の最大値は，頻出する語句として，「強く」，「風」と「雨」が現れており，風雨が強くなってきたことを示すトピックを示すツイートがまとまったクラスタあることが分かった．また，クラスタ 12 には，247 件のツイートが含まれており，頻出する語句として，「強く」，「風」，「雨」だけでなく，「晴れ」が含まれており，風雨が過ぎて，晴れてきていることを示すトピックを示すツイートがまとまったクラスタであることが分かった．

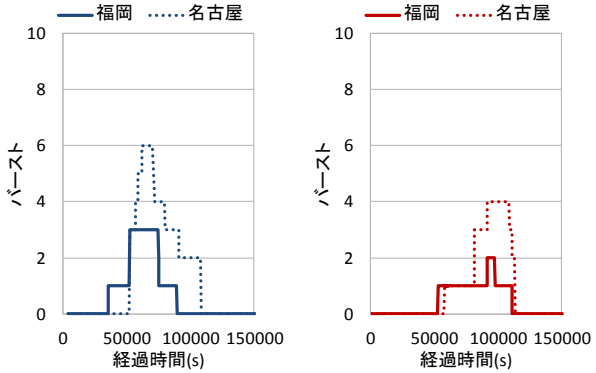


(a)東京　(b)名古屋



(c)大阪　(d)福岡

図 4.　バースト検出結果



(a)クラスタ 9　(b)クラスタ 12

図 5.　バーストの地域差

図 4 にバースト検出結果を示している．いずれのユーザの位置についても，クラスタ 9 がバーストして，次に，クラスタ 12 がバーストしていることが分かる．これは，台風が近づき，遠ざかることからもトピックに関するバーストを検出することができているといえる．また，台風の影響が強くなる時間と弱くなる時間とに，バーストの高い値が分かれており，クラスタのトピックがバーストの値としてグラフに表れていることが分かった．図 5 に，クラスタ 9 とクラスタ 12 の福岡と名古屋でのバースト状態の差を示す．台風が最初に近づいた福岡の方がバーストの立ち上がりが早く，あとから近づいた名古屋の方がバースト状態は長く続いていることが分かる．

V.　まとめ

本研究では，文書データをファジイ c-means 法により分類し，クラスタ毎にバーストを検出する手法を提案した．文書データをファジイ c-means 法により分類することでトピック毎にバーストが検出可能となった．また，クラスタへの帰属度とユーザとの距離を用いて文書データの影響度と定義した．提案手法では，影響度により Kleinberg のバースト検出手法で検出されるバーストに対して補正を行う手法となっている．評価実験の結果，実験に用いたデータにおいて，トピック毎のバーストを検出できていることを確認できた．これからの課題として，大規模データを用いて評価を行うこと，抽出したクラスタの定量的な評価，クラスタのトピック推定などがあげられる．

謝辞

本研究の一部は，広島市立大学・特定研究費（一般研究，研究課題名「時空間文書ストリーム上におけるバースト領域の抽出手法」） の支援により行われた．

参考文献

[1] J. Kleinberg, “ Bursty and Hierarchical Structure in Streams,” in Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 91–101, 2002.

[2] M. Naaman, “Geographic information from georeferenced social media data,” SIGSPATIAL Special, vol. 3, pp. 54–61, July 2011.

[3] Keiichi Tamura, and Hajime Kitakami, “Location-Based Burst Detection Algorithm in Spatiotemporal Document Stream,” in Proceedings of The 2012 International Conference on Data Mining (DMIN12), pp.195-201, July 2012.

[4] James C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms. Kluwer Academic Publishers, 1981.

問い合わせ先
〒731-3194
広島市安佐南区大塚東 3-4-1
広島市立大学情報科学部
事崎　翔太