画像付き文書データストリームにおけるバースト検出手法
Burst Detection Method for Image Document Stream

田村　真吾*　田村　慶一**　北上　始**　平原　海詞*
Shingo Tamura　Keiichi Tamura　Hajime Kitakami　Kaishi Hirahara
広島市立大学大学院情報科学研究科
Email:*{ mw67022, mu67026}@edu.ipc.hiroshima-cu.ac.jp, **{ktamura, kitakami}@hiroshima-cu.ac.jp

Abstract—**Extracting useful knowledge from a large-scale set of Web images, which are posted on the Internet, through social media sites, has become a new type of challenge. The main objective of this study is to extract the events and track the topics of a document stream that includes Web images, called an image document stream. This paper proposes a novel method for burst detection for an image document stream. The proposed method integrates a clustering technique with Kleinberg's burst detection. The experimental results show that the proposed method can extract the events and track the topics related to Web images posted on social media sites.**

I.　はじめに

近年，Flickr，Twitter や Facebook などのソーシャルメディアサイト上において，様々な時間や場所で撮影した写真を画像データとして投稿することで情報発信することが盛んに行われている．投稿される画像データは個人的な趣味だけでなく社会的な話題やイベントを含むため，ソーシャルメディアサイト上で投稿される画像データを分析する研究，イベントや話題などを取り出す手法の研究が行われている[1][2].

本研究ではソーシャルメディアサイト上で投稿される画像データを Web 画像データと呼ぶ．また，Web 画像は，通常，その Web 画像の内容を記載した文書データとともに投稿される．本論文では，この文書データのことを Web 画像付き文書データと呼ぶ．Web 画像付き文書データは時間とともに次々と投稿されるため，ストリームデータと捉えることができ，我々はこのストリームデータを画像付き文書ストリームとして定義する．

本論文では，画像付き文書ストリームからバーストを検出することによってイベントや話題などを取り出す手法を提案する．バーストとはストリームデータ上に現れるある事象が平常時と比較して頻繁に現れている状態を示す．例えば，ある観光スポットが話題になった場合，観光スポットの内容や様子とともに写真が Web 画像付き文書データとして投稿される．この投稿行動をバーストとして取り出すことができれば，観光情報としてユーザに有益な情報を提供することができる．

具体的には，提案手法では，テキストデータ内に記載される内容で Web 画像データを分類し，分類した Web 画像データ群（つまりクラスタ）ごとにバーストを検出する．クラスタリング手法として，特徴ベクトルベースの文書クラスタリング手法とグラフ分割ベースの文書クラスタリング手法の 2 つを検討する．また，バースト検出には，バースト検出アルゴリズムとして広い分野で利用や応用がされている Kleinberg のバースト検出アルゴリズム[3]を用いる．

提案手法を評価するために，画像 URL リンク付きのツイート 10,022 件を使用して評価実験を行った．クラスタリング手法として特徴ベクトルを使用した場合ではクラスタへの所属度が高い Web 画像付き文書データが集まった場合，関連するトピックを抽出できることを確認できた．一方，グラフ分割を使用した場合は，複数の細かいトピックがひつつのクラスタに集まり，取り出したトピックが大まかで曖昧な結果となった．

本論文の構成は以下の通りである．第 2 章では，バーストと Kleinberg のバースト検出手法について説明する．第 3 章では画像付き文書データストリームの定義と提案手法について述べる．第 4 章では評価実験結果を示し，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．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ことをバーストと呼ぶ.

Kleinberg のバースト検出手法では，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．

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

ここで，この式の第一項は，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．

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

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．

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

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．

$$\lambda_k = \frac{n}{T}\beta^k.$$

III. 提案手法

本章では，画像付き文書データストリームのデータモデルと提案手法について説明する．

A. データモデル

本研究では図 1 に示す画像付き文書データストリームを研究対象とする．図 1 は n 個の Web 画像付き文書データから構成される画像付き文書ストリーム $WIDS=\{wid_1,wid_2,...,wid_n\}$を示している．Web 画像付き文書データ d_i はその文書が到着した時刻 $alvtime_i$，その文書に添付された画像データ $wimage_i$ (Web 画像データの URL)，テキストデータ $text_i$ の三つの要素から構成され，$wd_i=<alvtime_i,wimage_i,text_i>$と表す．

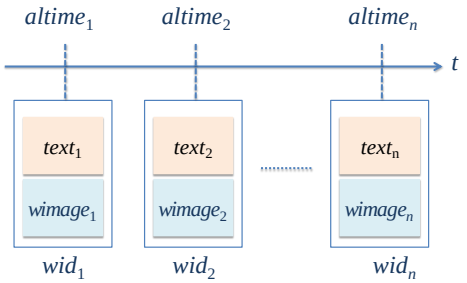


図 1. 画像付き文書データストリーム

B. クラスタリングベースのバースト検出

例えば，ある観光スポットが話題となると，観光スポットについて投稿される Web 画像付き文書データが増加していく．ただし，投稿される Web 画像データは観光スポットに関する同じ Web 画像データが投稿される場合（例えば，Twitter のリツートなど）だけでなく，別々の Web 画像データとともに投稿される可能が高い．観光スポットの Web 画像データに関するバースト検出する場合は，その観光スポットを扱う Web 画像データ群をひとつの単位としてバーストを検出する必要がある．

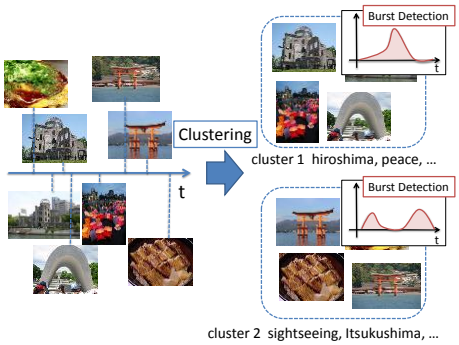


図 2. 提案手法

そこで，提案手法では，（1）Web 画像付き文書データのテキストデータを使用し，Web 画像データをクラスタリング手法で分類する．そして，（2）クラスタ毎にクラスタ内含まれる Web 画像付き文書データの到着間隔を元に Kleinberg のバースト検出アルゴリズムを使用してバーストを検出する．

図 2 に例を示す．図 2 の例では Web 画像付き文書データをクラスタリングすることで，広島，平和という話題を扱うクラスタ，観光，厳島という話題を扱うクラスタの 2 つのクラスタに分割している．そして，クラスタ毎にクラスタ内に含まれる Web 画像付き文書データの到着間隔を元にバーストを検出している．

C. アルゴリズム

テキストデータに含まれる語句を $T=\{term_1, term_2,...term_m\}$とし，以下に示す語句文書行列 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．

$$w_{i,j} = \begin{cases} \alpha_i & \text{if } term_i \text{ is included } doc_j \\ 0 \text{ if } term_i \text{ is not included } doc_j \end{cases}$$

特徴ベクトルベースのクラスタリングでは，語句文書行列の各列を Web 画像付き文書データの特徴ベクトルとして文書データを k 個のクラスタ Ci $(1 \leqq i \leqq k)$にクラスタリングする．

$$C_i = \{c_{i,j} | c_{i,j} \in WIDS, 1 \leq j \leq |C_i|\}$$

一方，グラフ分割ベースのクラスタリングでは，Web 画像付き文書データを頂点として表し，Web 画像付き文書データ間に共通する語句の割合が一定数以上の頂点を重み付き辺で接続した重み付き無向グラフを作成する．そして，重み付き無向グラフを Modularity ベースのグラフ分割手法を使ってクラスタに分割する．

ここで，各クラスタに含まれる Web 画像データ付き文書データの到着時刻のみを取り出して到着時刻ごとに並べ変えた系列を求める．この到着時刻系列は，

$$CALT_i = \left(calt_{i,1}, calt_{i,2}, \ldots, calt_{i,|c_i|}\right), calt_{i,j} \in ALT \quad ,$$

と表記する．ただし，*ALT* は *ALT*={$altime_1$, $altime_2$, ..., $altime_n$}と，すべての到着時刻の集合である．また，各クラスタに含まれる Web 画像データの集合を，

$$CWIMG_i = \left\{ cwimg_{i,j} \middle| \begin{matrix} cwimg_{i,j} \text{ is in } c_{i,j}, \\ cwimg_{i,j} \in WIMG, 1 \le j \le |C_i| \end{matrix} \right\}$$

と表記する．ただし，*WIMG* は *WIMG*={$wimage_1$, $wimage_2$, ..., $wimage_n$}と，すべての Web 画像データの集合を示す．

ここで，各クラスタの到着間隔の系列を$IAL_{C_i} = (ial_{i,1}, ial_{i,2}, \ldots, ial_{i,|C_i|})$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．

$$ial_{i,j} = \begin{cases} calt_{i,j+1} - stime, & j = 1, \\ calt_{i,j+1} - calt_{i,j}, & otherwise. \end{cases}$$

次に，各クラスタの各クラスタの到着間隔の系列を$IAL_{C_i} = (ial_{i,1}, ial_{i,2}, \ldots, ial_{i,|C_i|})$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．

$$C\left(s \middle| IAL_{C_i}\right) = \left(\sum_{i=1}^{n-1} \tau(s_i, s_{i+1})\right) + \left(\sum_{i=2}^{n} -\ln f_{s_i}(ial_{i,j})\right).$$

IV. 評価実験

本章では，評価実験の結果を示す．

A. データセット

実験で使用したデータセットは ANPI NLP で提供されている「東日本大震災」関連のハッシュタグを含む 3 月 11 日 15 時 16 分 9 秒から 25 日 8 時 59 分 19 秒までの画像 URL を含む 11,022 件のツイートである．

B. 実験条件

特徴量ベースのクラスタリング手法として Repeated-Bisection 法 [4] を用いた．Repeated-Bisection 法では，クラスタリングでは， 800 個のクラスタに分割した． Modularity ベースのグラフ分割手法として Newman 法[5]を用いた．Kleinberg のバースト検出におけるパラメータは，それぞれ，β=1.1，1.5， γ=0.1，0.01 である．

表 1. クラスタのランキング結果

順位	Web 画像付き文書データ件数	Web 画像データの種類	頻出語句
1	1081	1	規制，地震，大阪，条例
2	834	1	エネルギー，地震，余震，放出，沈静
3	386	1	一番，地震，声
4	282	1	原発，復興，自衛隊，被災，地震
5	214	2	地震，速報，緊急，楽譜
10	122	9	東北，支援，日本，メッセージ，太平洋
11	109	105	地震，確実，壁，ゴルゴ，酔い
16	87	2	地震，ワンピース，トイプードル，保護，首輪

C. 特徴量ベースの実験結果

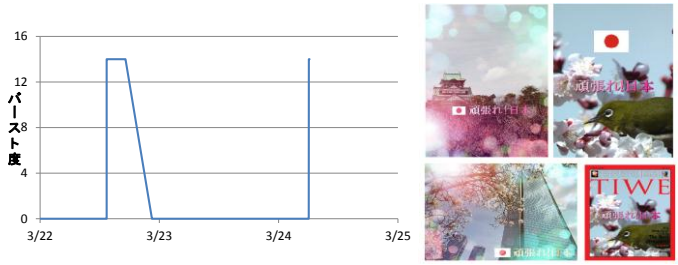
表 1 に，抽出したクラスタをクラスタに含まれる Web 画像データ数でランキングした上位 16 件（6 位～9 位，12 位～15 位は省略）を示す．表 1 では Web 画像付き文書データの件数だけでなく，そのクラスタに含まれる Web 画像データの種類を示している．

表 1 から，クラスタに所属する Web 画像データの件数は多いが，ほとんどが同じ Web 画像データであることが分かる．これは，震災という緊急的な内容であるためリツートの回数が多かったためだと考えられる．また，順位 1 と順位 3 は震災とは直接関係のない内容だが，文字では書き起こせない画像データで伝達するトピックを含むクラスタである．

図 3 に順位 10 位のクラスタのバーストとクラスタに含まれる Web 画像データの一部を示す．このクラスタに含まれる画像は「ガンバレ日本！」と書かれた似たテーマを持つ．クラスタに所属する個々の Web 画像データの発生回数でランキングするとそれぞれは下位になるが，クラスタリングを行うことで同じトピックの Web 画像データをまとめランキングを上位にすることができる．また，バーストは期間の後半から始まっており，復興を目指すようになった時期からこのような Web 画像データが投稿され始めたことを示している．

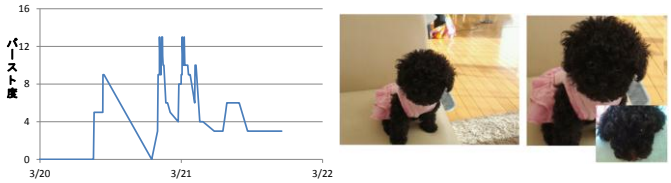
図 4 に順位 16 位のクラスタは地震の際に行方不明になった愛犬を探してほしいという旨の Web 画像データを含み，ある特定の区間でバー

ストしている．画像データの発生回数の順位が低い画像もクラスタとしてまとまっているために順位を上げることができているといえる．



(a)バーストの変化　(b)含まれていた画像群

図 3. 順位 10 位のクラスタ



(a)バーストの変化　　(b)含まれていた画像群

図 4. 順位 16 位のクラスタ

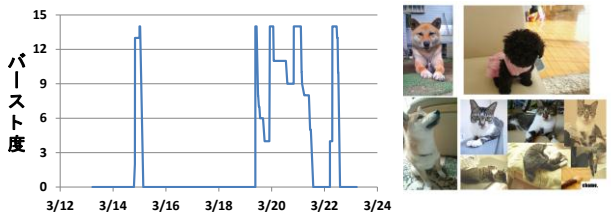
D. グラフ分割ベースの実験結果

グラフ分割ベースのクラスタリングでは 387 個のクラスタが得られた．特徴的なクラスタを図 5 に示す．クラスタには計 313 個の画像が含まれておりどれもツイートの内容は行方不明の自分のペットを探してほしいというものであった．キーワードにも「首輪」，「行方」，「不明」，「お願い」と言ったものが多くペットの捜索に関するトピックであるといえる．

しかしながら，図 6 に示すように，トピックの検出が難しいクラスタも見受けられた．図 6 のクラスタには計 1913 枚の画像が含まれていたが，複数種類のキーワード群や画像群が存在し，一つのクラスタにトピックが複数含まれている状態となった．その為，バーストも散逸的である一つの決まったトピックを抽出することが出来なかった．

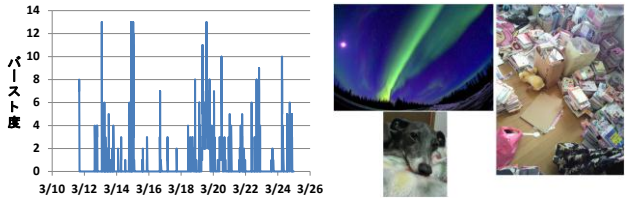
V.　まとめ

本論文では，Web 画像付き文書データに含まれる文書データに着目し，画像付き文書データストリーム上の Web 画像付き文書データをクラスタリング手法で分類し，同じ話題を扱う画像データ群や類似した画像データ群を一つの単位としてバーストを検出する手法を提案した．クラスタに分類することでイベントや話題単位でバーストを検出することができた．一方，クラスタリングがうまくいっていないクラスタはクラスタとして統一したイベントや話題を扱うことができなかった．これからの課題として，クラスタリングの精度向上があげられる．



(a)バーストの変化 (b)含まれていた画像群

図 5.グラフ分割ベースで得られたクラスタ 1



(a)バーストの変化(b)含まれていた画像群

図 6. グラフ分割ベースで得られたクラスタ 2

謝辞

本研究の一部は，文部科学省・科学研究費補助金(若手研究(B)，課題番号：23700124)，日本学術振興会・科学研究費補助金(基盤研究(C)，課題番号：20500137)の支援により行われた．

参考文献

[1] N. A. Van House, “Flickr and public image-sharing: distant closeness and photo exhibition,” in CHI ’07 extended abstracts on Human factors in computing systems, CHI EA ’07, pp. 2717–2722, 2007.

[2] V. K. Singh, M. Gao, and R. Jain, “Social pixels: genesis and evalu¬ation,” in Proceedings of the international conference on Multimedia, MM ’10, pp. 481–490, 2010.

[3] J. Kleinberg, “Bursty and hierarchical structure in streams,” in Pro¬ceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD ’02, pp. 91–101, 2002.

[4] George Karypis, Eui-Hong Han, Vipin Kumar, “CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling,” in IEEE Computer 32(8), pp. 68-75, 1999.

[5] NEWMAN M. E. J., “The structure and function of complex networks,” SIAM Review 45, pp.167-256, 2003.

問い合わせ先

〒731-3194

広島市安佐南区大塚東 3 丁目 4 番 1 号

広島市立大学　情報科学研究科　知能工学専攻

田村　真吾