画像付きツイートに対するキーワード検索結果の要約手法
Summarizing the Results of a Keyword Search on Tweets with Web-Images

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

Abstract— **Recently, more and more people post tweets with Web-image links in order to spread various types of applications cooperated with camera on Twitter site. In the last few years, detecting extracting useful knowledge in a set of tweets with Web-image links has attracted much attention of researchers, because it contributes to marketing, tourism informatics, and recommendation systems. The main objective of this study is to develop a novel summarization method for keyword search on tweets with Web-Image links. A huge number of results are returned from a keyword search on tweets with Web-Image links. It is difficult for user to understand the results of the keyword search. Although, there are many conventional summarization techniques for text data, the conventional summarization techniques cannot make summary that take account time changes. To overcome this difficultly, we propose a novel summarization method for keyword search on tweets with Web-Image links using the clustering-based burst detection algorithm. The experimental results show that the proposed method can extract summaries that are included the bursts of topics and events taking account time changes.**

I.　はじめに

カメラ付きスマートフォンの普及とともに，ソーシャルメディアサイト上において，ユーザは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を行うようになってきている．ソーシャルメディアサイト上において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み，一種の集合知を形成し始めている．そこで，投稿される画像データの分析，観光情報への応用，イベントや話題などを取り出す手法の研究が盛んに行われている[1][2].

本研究では，ソーシャルメディアサイトのひとつである Twitter のツイートとともに投稿される画像データに着目し，画像データ（画像データの URL）付きで投稿されるツイートを画像付きツイートと呼ぶ．画像付きツイートは，通常，画像データの内容を記載した文書データとともに投稿される．例えば，あるイベントに参加したユーザはイベント会場を撮影した画像データとともに，イベント名とコメントを投稿し，情報発信を行うことが考えられる．

画像付きツイート集合に対してキーワード検索を行うことで，関心のある投稿画像データを検索することができる．しかしながら，キーワード検索結果として大量のツイートが得られても，検索結果を時系列に閲覧することで，検索結果に含まれるトピックやイベントを手作業で探し出す必要がある．画像データの投稿数で注目されている画像データを表示するサービスも提供されているが，時間的な変化として注目度を取り出すことができない．

そこで，本研究では，画像付きツイート集合に対するキーワード検索結果に含まれるトピックやイベントをユーザに時間変化の要約という分かりやすい形で提示するための手法として，クラスタリングに基づくバースト検出アルゴリズム[3]を応用することを試みる．クラスタリングに基づくバースト検出アルゴリズムを用いることで，関連が高い画像データをまとめるだけでなく，注目度の時間的な変化をバーストという形で要約してユーザに提示することができる．

本論文の構成は以下の通りである．第 2 章では，クラスタリングに基づくバースト検出アルゴリズムで用いる Kleinberg のバースト検出アルゴリズムについて説明する．第 3 章では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.　KLEINBERG のバースト検出アルゴリズム

本章では，クラスタリングに基づくバースト検出アルゴリズムの説明に必要となる，文書ストリーム，バーストと Kleinberg のバースト検出アルゴリズム[4]について説明する．

文書ストリームは，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} が到着するという時系列の文書データ系列である．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．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

タは次々と投稿される．そして，投稿される文書データ間の到着間隔は短くなるため，到着間隔に着目するとバーストを検出することができる．

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．

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

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，データモデル，クラスタリングに基づくバースト検出手法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 に示す画像付きツイートを Web 画像付き文書データストリームとして扱う．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，wid_i=<$altime_i$,$wimage_i$,$text_i$>と表す．

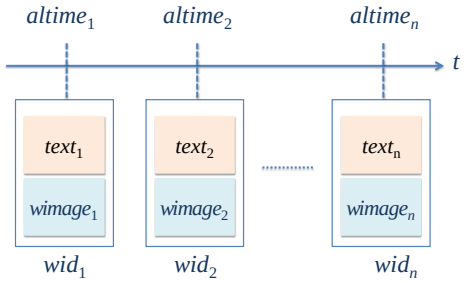

図 1. Web 画像付き文書データストリーム

B. クラスタリングに基づくバースト検出

クラスタリングに基づくバースト検出アルゴリズムでは，（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．

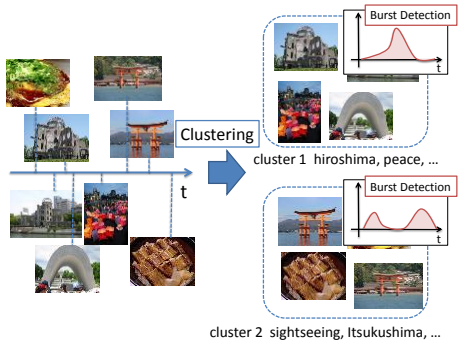

図 2. クラスタリングに基づくバースト検出

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. 要約アルゴリズム

キーワード検索のキーワードを $keyword$ とする．Web 画像付き文書データストリーム $WIDS$ において，キーワード $keyword$ をテキストデータに含む部分ストリーム空間 $WIDS^{keyword}$ とする．

$$WIDS^{keyword}=\{wid_{\phi(1)},wid_{\phi(2)},...,wid_{\phi(l)}\}$$

ただし，ϕは，$WIDS^{keyword}$ から $WIDS$ への単射で，キーワード $keyword$ をテキストデータに含む Web 画像付き文書データの数を l とする．ここで，$WIDS^{keyword}$ の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 $W^{keyword}$ を作成する．語句文書行列 $W^{keyword}$ 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．

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

語句文書行列$W^{keyword}$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 C_i $(1 \leqq i \leqq k)$にクラスタリングする.

$$C_i = \{wid_{\omega_i(1)}, wid_{\omega_i(2)}, \ldots, wid_{\omega f_i(numd(i))}\}.$$

ここで, ω_iはC_iを構成する文書データ集合から*WIDS*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. 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,

$$CALT_i = (altime_{\omega_i(1)}, altime_{\omega_i(2)}, \ldots, altime_{\omega_i(numd(i))}).$$

と表記する. また, 各クラスタに含まれる Web 画像データの系列は,

$$CWIMG_i = (wimage_{\omega_i(1)}, wimage_{\omega_i(2)}, \ldots, wimage_{\omega_i(numd(i))})$$

となる.

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.

$$ial_{i,j} = \begin{cases} alvtime_{f_i(j)} - stime, & j = 1, \\ alvtime_{\omega f_i(j)} - alvtime_{\omega_i(j-1)}, & otherwise. \end{cases}$$

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.

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

コスト式を最小化する状態遷移系列は, ビタビアルゴリズムを用いて求めることができる. 最後に, 画像データ数でクラスタをランキグし, クラスタ毎の状態遷移列, 頻出語句, 頻出画像データをユーザに返却する.

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

Twitter Streaming API により取得した, 地域を「日本」であると申請しているユーザがつぶやいたツイートの中から, Twitter 公式 (pic.tiwtter), Twitpic と Yfrog の URL を含むを平成 23 年 11 月 14 日の 5 : 46 : 51～同月 24 日 00 : 55 : 29 の間のツイート 380 万件を対象とした.

B. 実験条件

キーワード「雪」と「映画」を用いてキーワード検索を行った結果の要約結果の評価を行った. クラスタリングに基づくバースト検出手法では, 特徴量ベースのクラスタリング手法として Repeated-Bisection 法[5]を用いて, クラスタリングを行った. そして, 上位 5 件のクラスタから得られるバースト検出結果, 頻出語句, 頻出画像データを要約として取り出した. クラスタリングのクラスタ数は 50 個, β 値＝ 1.1, 1.5, γ 値＝ 0.1, 0.01 の組み合わせ計 4 通りのバースト性パラメータを用いて評価を行った.

C. 実験結果

表 1 にキーワード「雪」, 表 2 にキーワード「映画」の要約結果として得られた上位 5 件のクラスタを示す.

表 I. 「雪」の要約結果

順位	画像データ数	頻出語句
1	7706	雪, 歩, ちょっと, こんな, かんじ
2	1295	積雪, 吹雪, 雪国, センチ, 札幌
3	1058	初雪, 冬, 今日, わさ, おはようございます
4	1023	雪だるま, カール, 大阪, 駅, かわいい
5	937	雪, 今日, 日, おはようございます, wbs

表 II. 「映画」の要約結果

順位	画像データ数	頻出語句
1	2569	映画, 公開, けいおん, 監督, 怪物
2	443	映画館, これ, 今日, 映像, 上映
3	356	投票, 人気, 映画, イナズマ, 順位
4	305	映画, 明日, 発売, 公開, 怪物君くん
5	293	映画, 今日, これ, 笑, みたい

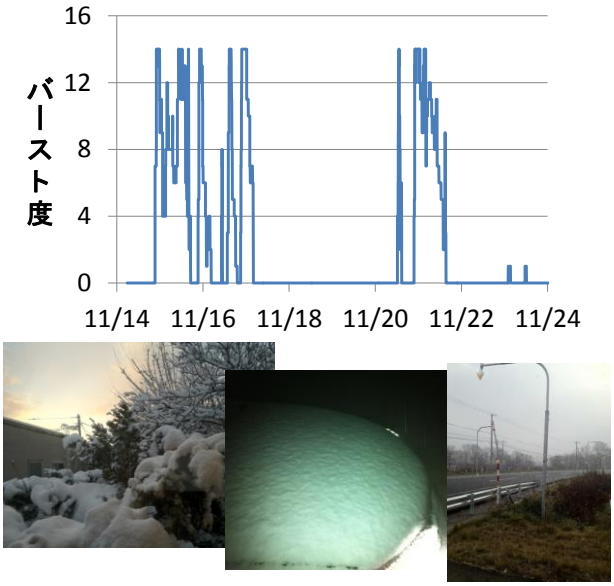

図 3.　検索ワード「雪」におけるクラスタ (1 位) のバーストと頻出画像

図 3 に「雪」における 1 位のクラスタの要約結果を示す. このクラスタは全国的に北日本で初雪が降ったことを示すクラスタで, 初雪が降った期間がバーストとして抽出されている. 図 4 に「映画」における 1 位のクラスタの要約を示す. 「けいおん」という単語が最も多く画像も「けいおん劇場版」に関する画像が多かった. 公開日の 11 月 15 日から頻繁にバーストしており, 本作の人気が垣間見える. ただし, 映画を日中見る人がツイートをするためバーストはこの期間, 毎日現れており, 時間的な特徴を取り出せているとはいえなかった.

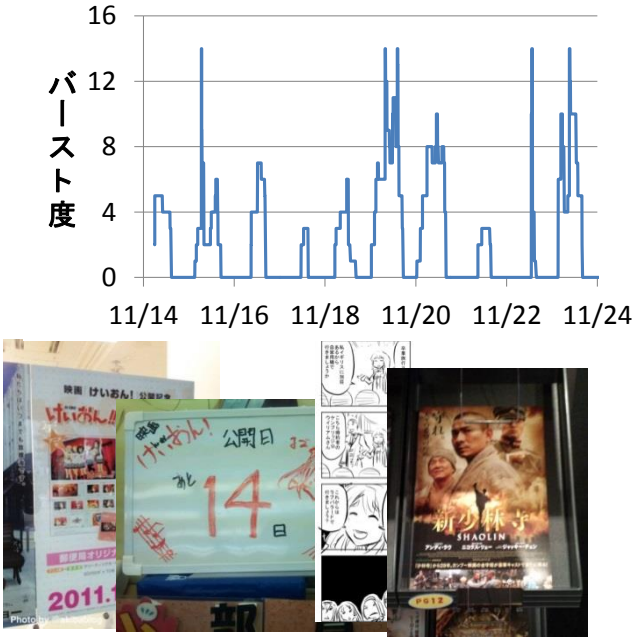

図 4.　検索ワード「映画」におけるクラスタ（1 位）のバーストと頻出画像

V.　まとめ

本研究では, 画像付きツイート集合に対するキーワード検索結果に含まれるトピックやイベントをユーザに時間変化の要約という分かりやすい形で提示するための手法として, クラスタリングに基づくバースト検出アルゴリズムを応用することを試みた. クラスタリングに基づくバースト検出アルゴリズムを用いることで, 関連が高い画像データをまとめるだけでなく, 注目度の時間的な変化をバーストという形で要約してユーザに提示することができる. しかし, 検索する単語により精度がまちまちであり, この点の結果をどう扱うかを考察する必要がある. これからの課題として, 各クラスタがどのようなトピックやイベントを扱っているのかを自動判定する手法, クラスタを代表する画像データの抽出, テキストデータと画像データの両方を使ったクラスタリングなどがあげられる.

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] 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] Shingo Tamura, Keiichi Tamura, Hajime Kitakami , and Kaishi Hirahara, “Clustering-based Burst-detection Algorithm for Web-image Document Stream on Social Media,” in Proceedings of IEEE SMC 2012, pp.703-708, 2012.

[4] 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.

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

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