

土産に関するツイートのバースト現象の分析

Analysis of Burst of Tweets Containing Souvenir Information

真辺諒*¹
Ryo Manabe

長尾哲志*¹
Noriyuki Nagao

安藤一秋*²
Kazuaki Ando

*¹ 香川大学大学院工学研究科
Graduate School of Engineering, Kagawa University

*² 香川大学工学部
Faculty of Engineering, Kagawa University

The aim of our research is to construct a system to provide users useful information about souvenirs that can purchase only at the particular location or area by extracting the information from blog sites and question and answer sites. In order to confirm the usefulness of Twitter as a resource of real-time information about souvenirs, we analyze topic words and tweets extracted from burst of tweets.

1. はじめに

旅行に行った際、多くの人が土産を購入するが、土産の選定は悩ましい問題である。土産の選定時に、現地でしか購入できない商品や定番商品、現在注目されている商品およびそれらの口コミ情報などがわかれば、この状況を改善できると考える。

そのような中、長尾ら[長尾 2015]は、現地でしか購入できない土産に関する情報をQAサイトや口コミサイト、ブログから自動収集して、ユーザに提示するシステムを提案した。しかし、長尾らが対象とする土産情報は、比較的静的な情報であるため、リアルタイム性や話題性に関する情報が不足する。

本稿では、長尾らのシステムにおいて、リアルタイム性や話題性など、土産に関する動的な情報を補うための情報源として、Twitter に注目する。Twitter は即時性や簡便性の高さから多くの人が利用し、日々大量の情報を発信している。現在、月間アクティブユーザ数は3億1千万人、一日に投稿されるツイート数は5億件を超える。国内ユーザ数は2016年9月の時点で4千万人となっており、およそ日本人の3人に1人が利用していることになる。このように日々大量に発信される情報の中には、土産に関する情報も存在している可能性がある。また、Twitter では、ある話題に対するツイートが急上昇し、平常時と比べてツイート数が大きく増加する「バースト」と呼ばれる現象が観測される。土産に関するツイートのバーストを検出することで、注目されている商品名や店舗名、レビューなど、土産の選定時に有用な情報を取得できる可能性がある。

そこで本稿では、Twitter から土産に関するツイートを収集し、その集合内で発生するバーストを検出・分析し、土産情報を提示するシステムにおける Twitter の有用性を確認することを目的とする。具体的には、まず、土産に関するツイートを収集し、その集合内で発生するバーストを1日単位で検出する。次に、バーストが検出された日のツイート群から話題語を抽出し、それらの話題語をもとに注目ツイート(以降、トピックツイート)を抽出する。そして、抽出された話題語とトピックツイートを調査・分析することで、土産情報を提示するシステムにおける Twitter の有用性を考察する。

2. バーストの検出

2.1 土産に関するツイートの収集

バーストの検出に利用する土産に関するツイートの収集法について述べる。

本稿では、Twitter の Search API を利用して、キーワード「土産」または「みやげ」を含むツイートを収集する。2016年7月1日から2017年2月28日の約8ヶ月間に渡って収集した結果、土産に関する865,161件のツイートが収集できた。

以降、このツイート集合におけるバースト現象について調査・分析する。

2.2 バースト検出手法

水沼らの研究[水沼 2013]では、何らかの要因が影響することにより、平常時と比較してツイート数が大きく増加する現象を「バースト」と定義している。また、一定の閾値を定め、ツイート数が閾値を超えた場合をバーストとして検出する手法を提案している。本稿では、水沼らの手法を参考にバーストを検出する。

水沼らは、閾値の設定を1時間単位で行っているが、本研究では土産に関するツイート数を考慮し、1日単位で設定する。本研究での閾値は、1日ごとのツイート数の平均と標準偏差をもとに、以下の式で設定し、ツイート数が閾値を超えた日をバーストとして検出する。

$$N_{nd} = \bar{N} + \sigma \quad (1)$$

ここで、 N_{nd} は閾値、 $\bar{N}$ は1日あたりの平均ツイート数、 σ は標準偏差である。

バーストの検出にあたり、ツイート数の最大値、最小値は外れ値としてツイートデータから除外する。また、平均ツイート数に関して、水沼らは平日と休日とで投稿数の分布に差異が見られるとしている。本稿でも、平日と休日を分けて閾値を設定する。

2.3 バーストの検出結果

閾値 N_{nd} と1日ごとのツイート数の分布を比較し、バーストを検出した結果、平日163日において23日、休日80日において11日のバーストが検出された。図1に閾値と1日ごとのツイート数の分布の比較を示す。

図1を見ると、8月中のツイート数がかなり増加していることがわかる。また、11月の後半と年末年始にかけて、閾値を上回るツイート数の増加が見られることから、バーストが発生していることがわかる。

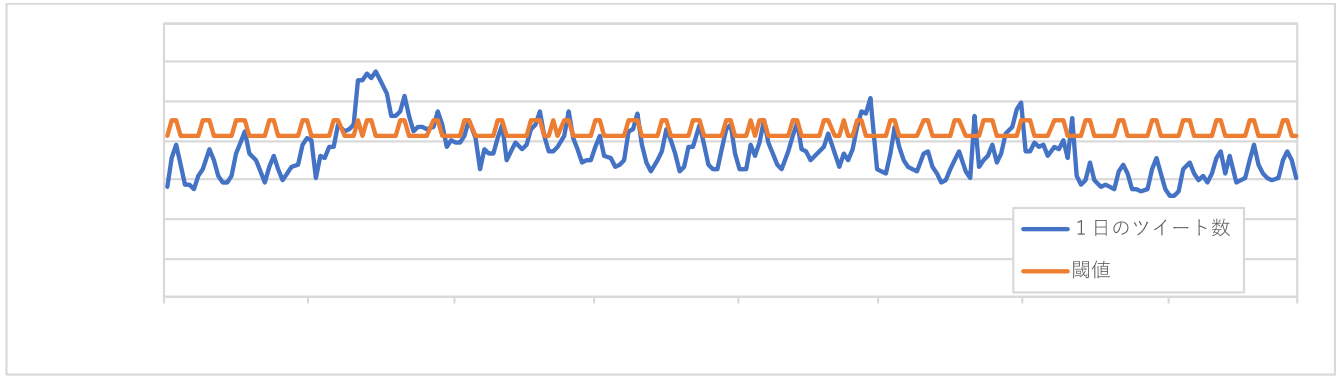


図 1. 閾値と1日ごとのツイート数の分布の比較

本稿では、バーストの検出日のうち、2016年8月15日(夏季)と2016年12月21日(冬季)のツイートを対象に、ツイート内に出現する話題語とトピックツイートを抽出し、調査・分析する。

3. 話題語とトピックツイートの抽出

バーストが検出されたツイート群から話題語を抽出する方法と、抽出した話題語をもとにトピックツイートを抽出する方法について述べる。

3.1 話題語の抽出

話題語の抽出には、藤川ら[藤川 2012]の手法を参考にする。藤川らは、単語の出現確率として二項分布を仮定している。まず、単語 w の平均出現確率 p を以下の式で求める。

$$p = \frac{w \text{ が出現したツイート数}}{\text{全ツイート数}} \quad (2)$$

そして、1日あたりのツイート n 件のうち、 k 件のツイートにおいて、 w が出現する確率 $prob$ を以下の式で求める。

$$prob(n, k) = p^k (1 - p)^{n-k} \binom{n}{k} \quad (3)$$

藤川らの研究を参考に、1日あたりのツイートにおいて、単語 w の出現確率が平均出現確率 p を上回り、(3)の式により算出した w の出現確率が 0.05 以下の場合、単語 w を話題語として抽出する。そして、抽出された話題語のスコア $TopicWord$ を以下の式で求める。

$$TopicWord(w) = -\log prob(n, k) idf(w) \quad (4)$$

ここで、 $idf(w)$ は、その月における単語 w の idf 値である。

3.2 トピックツイートの抽出

抽出した話題語 w をもとに、トピックツイートを抽出する。 i 番目のツイートに対し、トピックツイートとしての重要度 $TopicTweet$ は、以下の式で求める。

$$TopicTweet(i) = \sum_j^{n_i} TopicWord(w_{ij}) \times \frac{n_i}{m_i} \quad (5)$$

ここで、 n_i はツイート i に含まれる話題語の数、 m_i はツイート i に含まれる名詞の数である。

式(5)を用いて、各ツイートの重要度を計算し、降順にソートする。その内の上位 α 件をトピックツイートとする。

4. 話題語とトピックツイートの抽出と分析

バーストが検出された2016年8月15日(夏季)と2016年12月21日(冬季)のツイートを対象に、話題語とトピックツイートを抽出し、その特徴を分析する。

4.1 前処理

MeCab でツイートを形態素解析し、名詞のみを話題語として抽出する。このとき、不要語処理として、名詞のうち、「接尾」、「非自立」、「代名詞」、「ナイ形容詞語幹」を除去する。また、絵文字や記号、数字、2文字以下の英字も除去する。

URL 付きツイートには、外部サイトを宣伝するものが多く、類似した内容のツイートが多数つぶやかれた場合、重要なツイートが埋もれてしまう可能性がある。そのため、トピックツイートの重要度を計算する前段階で、類似している URL 付きツイートは1件のみを残し、その他は削除する。類似度の判定には、3-gram を利用し、類似度が 0.4 以上のツイートを削除する。また、ツイートの内容を完全に包含する他のツイートがある場合、後者を代表ツイートとする。

4.2 話題語の抽出結果

2016年8月15日(夏季)と2016年12月21日(冬季)のツイートから抽出した話題語を表1と表2にそれぞれ示す。表中の話題語は、スコア順における上位10件である。

表 1. 2016/8/15 のツイートから抽出した話題語(上位 10 件)

1	角煮まんじゅう	6	サクッ
2	ロッキン	7	夏コミ
3	上り	8	じゃらん
4	小松空港	9	石川
5	中央	10	プチプラ

表 2. 2016/12/21 のツイートから抽出した話題語(上位 10 件)

1	はてなブックマーク	6	硬貨
2	タカラトミー	7	節約料理
3	タカラトミーアーツ	8	寒梅
4	亀田製菓	9	ハッピーターン
5	付け所	10	コナコーヒー

表 3. 2016/8/15 のトピックツイート(上位 5 件)

1	中央道上り線で見つけて欲しい！会社を持って行って喜ばれる長野・山梨土産 #グルメ https://t.co/4dZTyDQC2C By ippin 編集部 八ヶ岳や富士五湖など高原リゾート地が多い中央道。夏に限らず連休になると混雑す… https://t.co/GRhVovJB6G
2	長崎土産 岩崎本舗の、角煮まんじゅう(へへ)
3	長崎土産はカステラ？ ではなく角煮まんじゅう、だって。
4	@xxxxx 土産やるわ ばいなら 明日から再びお盆休みへ
5	@yyyyyy_夏コミお疲れ様でした。 ガルパン本の新刊(個人、合同とも)も買わせていただきました。ただただカワイイ本で最高でした！ ガルパン本に合わせて大洗のガルパン土産も差し入れさせて頂きました。 また次回の本を楽しみにさせていただきます。

表 4. 2016/11/29 のトピックツイート(上位 5 件)

1	柿の種に越乃寒梅 ハッピーターンも 亀田製菓、土産商品拡充 - 日本経済新聞 https://t.co/e9vgZHVWAJ
2	ナイスな発想！成田空港に大量ガチャ 両替できなかった小銭が土産に変身 日本らしさ重視、売り上げ 3 倍強 https://t.co/03dCwZewNb これはいい発想だと思う。あと、余った硬貨はボランティアに寄付できるような仕組みもいいんじゃないのかな。
3	これは素晴らしい。硬貨を持って帰っても使い道はないです。ガチャなら日本土産にぴったりです。 ナイスな発想！成田空港に大量ガチャ 両替できなかった小銭が土産に変身 日本らしさ重視、売り上げ 3 倍強 https://t.co/WNwuwBqwGm #Yahoo ニュース
4	ナイスな発想！成田空港に大量ガチャ 両替できなかった小銭が土産に「あまった小銭をオモチャに！」を合言葉に、 外国人ウケしそうな 171 台を設置したのはタカラトミーアーツ。詳しく話を聞きました。...J... ★ https://t.co/scgRQnNtZA #おもちゃ
5	ナイスな発想！成田空港に大量ガチャ 両替できなかった小銭が土産に変身 日本らしさ重視、売り上げ 3 倍強 https://t.co/wrjMCbQ00H 確かに海外の人ってアキバとかでガチャに釘付けの光景よく見るし、両替できなかった小銭問題もある。発想の大勝利。

4.3 トピックツイートの抽出結果

4.2 で抽出した話題語を用いて、2016 年 8 月 15 日(夏季)と 2016 年 12 月 21 日(冬季)のツイートからトピックツイートを抽出した結果を表 3 と表 4 にそれぞれ示す。表中のトピックツイートは、重要度順における上位 5 件である。

4.4 話題語とトピックツイートの分析

抽出した話題語とトピックツイートについて分析する。まず、表 1 に示す 2016 年 8 月 15 日(夏季)のツイートから抽出された話題語をみると、商品名として「角煮まんじゅう」が 1 位で取得できていることがわかる。そして、表 3 のトピックツイートを見ると、「角煮まんじゅう」に関するツイートがトピックツイートとして 2 件収集できている。調査の結果、当日のテレビ番組で長崎県の角煮まんじゅうに関する内容が放送されていたことがわかった。表 3 の 2 位のトピックツイートには店舗名も記載されているため、有用なツイートである。

表 3 の 1 位のトピックツイートは、「長野・山梨土産」に関するもので、グルメ紹介サイトから発信されたツイートであった。「中央」や「上り」という話題語に関連し、また、10 位以降の話題語として「長野」や「山梨」のスコアも高かったため、このツイートの重要度が高くなったと考える。

表 3 の 4 位のトピックツイートに関しては、土産に関して有用な情報が含まれないが、重要なツイートであると判定された。これは、「お盆」と「休み」がスコアの高い話題語であったことが要因と考えられる。

また、表 1 の話題語「夏コミ」に関しては、夏のコミックマーケットの開催期間が近かったことが原因で多数ツイートされたと考え

られる。そのため、「夏コミ」に関連するツイートも重要度の高いツイートと判定された。

次に、表 2 に示す 2016 年 12 月 21 日(冬季)のツイートから抽出された話題語と表 4 に示すトピックツイートを見ると、成田空港にガチャガチャが設置された内容のツイートが 4 件を占めていることがわかる。また、メディアが取り上げたニュースに対して、意見・感想が併記されているものが多いことがわかる。これらのツイートは、話題語として、「タカラトミー」や「硬貨」の重要度が高いことが影響し、抽出されたと考える。

表 4 のトピックツイートの内、1 位と判定されたツイートに関しては、メディアが発信したツイートであるが、「亀田製菓」や「寒梅」、「ハッピーターン」など、商品名に関連する話題語が含まれていることから、重要度が高くなったと考える。

以上より、2 日分の調査結果ではあるが、土産に関するツイート群で発生したバーストから、一部、土産に関する有用な情報を収集できるが、メディアで取り上げられた土産やサイトでの紹介が多くを占める傾向を確認した。したがって、土産に関するツイートから商品名や店舗名、レビューなどの有用な情報を大量に得ることは難しいが、メディア上で取り上げられた土産に関する話題について、盛り上がり度や注目度などとして、活用できる可能性があると考えられる。

5. おわりに

本稿では、土産情報提示システムに対し、リアルタイム性や話題性など、土産に関する動的な情報を補うため、ツイートのバーストに注目し、話題語とトピックツイートの観点で分析した。

まず、土産に関する 8 か月分のツイートから、平日 163 日において 23 日、休日 80 日において 11 回のバーストが検出され

た.そして、話題語とトピックツイートの抽出結果から、メディアで取り上げられた土産に関するツイートが多数を占める可能性を確認した.これらの結果から、メディア上で取り上げられた土産に関する話題について、盛り上がり度や注目度などの情報をTwitterから抽出し、活用できる可能性があると考え.

今後の課題として、話題語のスコア調整とクラスタリング、土産に関する注目度・話題度の可視化、新商品に関するツイートの抽出などを検討する.

参考文献

- [長尾 2015] 長尾哲志, 安藤一秋:オンラインショップで購入できない土産を提示するシステムの構築, 第14回情報科学技術フォーラム講演論文集, pp.71-72, 2015.
- [水沼 2013] 水沼友宏, 池内淳, 山本修平, 山口裕太郎, 佐藤哲司, 島田諭:Twitterにおけるバーストの生起要因と類型化に関する分析, 情報社会学会誌, Vol.7, No.2, pp.41-50, 2013.
- [藤川 2012] 藤川智英, 鍛冶伸裕, 吉永直樹, 喜連川優:マイクログログ上の話題抽出とユーザの態度の分類に基づく流言検出支援システム, 第4回データ工学と情報マネジメントに関するフォーラム(第10回日本データベース学会年次大会), F7-1, 2012.
- [白川 2010] 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, 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, Vol. 2010-DBS-150, No.2, pp.1-8, 2010 年.
- [前川 2013] 前川 浩基, 中原孝信, 羽室行信:テレビ番組視聴時における twitter 投稿のバースト検知と情報配信の可能性, 第27回人工知能学会全国大会論文集, pp.1-4, 2013.