クエリログを用いた時空間分析に基づく地域特性の抽出

佐久間 幸太郎† 北山 大輔‡ 角谷 和俊†
†関西学院大学総合政策学部 〒669-1377 兵庫県三田市学園２丁目１
‡工学院大学情報学部 〒163-8677 東京都新宿区西新宿１丁目２４番２号
E-Mail：†{dxv49790, sumiya}@kwansei.ac.jp ‡kitayama@cc.kogakuin.ac.jp

あらまし

本研究は，地域特性の抽出のための手法の提案を行う．ある地域における特徴的な事柄の検索数は，他の地域におけるその事柄の検索数の推移や周期性などの時系列特徴に違いがあると考えられる．本論文ではこの時系列特徴の違いに注目し，地域特性の抽出を行う．時系列特徴の違いの抽出方法として，各地域の時系列特徴から一般的（全国的）な時系列特徴を差し引いて求めた値（差異）を利用する．この差異の値を利用し，複数の地域間における相関係数を算出した．算出された結果から他の地域との相関が低い，または負の相関をもつ地域では，その事柄に関する地域特性をもつと考えられるため，それらの地域から地域特性となるパターンを抽出した．また，他のキーワードと比較することで，地域特性に関する特徴語の抽出を図る．

キーワード：クエリログ，時空間分析，時区間分析，地理情報，地域特性

１．はじめに

現在，毎年外国人観光客が増え続け，2015 年には約 1974 万人の外国人観光客が来日した[1]．2020 年に東京五輪が開かれることになり，ますますの外国人観光客の増加が見込まれる．また観光庁のデータ[2]によると，近年では日本の四季にふれた観光が盛んとなってきており，桜や紅葉をはじめとする一定の時期にしか体験することのできない観光の形式が増えてきている．

本研究ではこのようなある一定の時期における観光の促進のため，クエリログを用いた時空間分析に基づく地域特性の抽出のための手法を提案する．特徴的な事柄（地域特性）がある地域においては，それの検索数に他地域とは異なる時系列特徴がみられると考えられる．この仮説から，時系列特徴を利用した地域特性の抽出を図る．地域特性の抽出の際，単なる検索数の時系列データを利用するのではなく，時系列データから一般的（全国的）な時系列データを差し引いた値（差異）を利用する．こうすることにより，一般的特徴を差し引いた地域特性の抽出を行う．

本論文では，2 章で概要を述べたあと，3 章で地域と全国との違いを利用した分析方法を提案する．また，相関係数を利用して他地域との比較を行い，他の事柄との比較による地域特性に関する特徴語の抽出を行う．4 章では関連研究についてまとめ，最後に 5 章でまとめと今後の課題を述べる．

２．研究概要

本研究では，ある事柄における各地域間の検索数の違いを利用した地域特性抽出の一手法の提案を行う．本研究では Google Trend[3]から直近 5 年間（2012 年から 2016 年）の週別データを抽出し，各週の 5 年間平均の時系列データを用いた分析を行う．

ある地域における特徴的な事柄（＝地域特性）の検索数は，他の地域におけるその事柄の検索数と時系列特徴に大きな違いが存在すると考えられる．これらの時系列特徴の地域間の違いを利用し，地域特性の抽出を図る．ここで，本研究における「地域特性」とは，ある観点からみたときにある地域に限定してみられる時系列特徴のことを指すものとする．例えば，一般的に「キャンプ」の検索数は図 1 のように，ゴールデンウィークや盆の時期に検索数のピークがみられる．それに対して，「宮崎 キャンプ」の検索数は図 2 のように，2 月に検索数のピークがみられるグラフとなる．この時系列特徴

は「キャンプ」の検索数の推移のグラフにはみられないため，「宮崎　キャンプ」における時系列特徴ということができる．

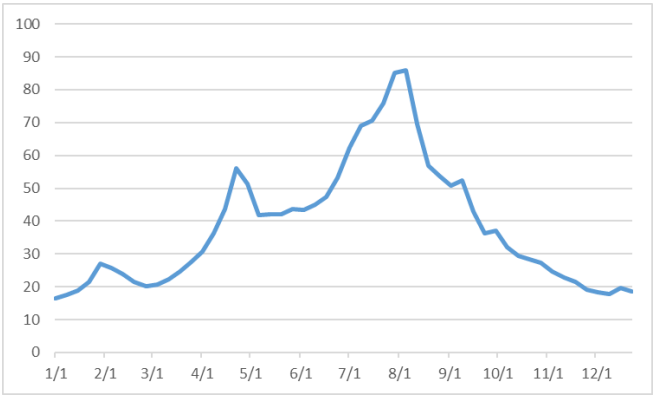

図 1「キャンプ」の検索数の推移（5 年間平均）

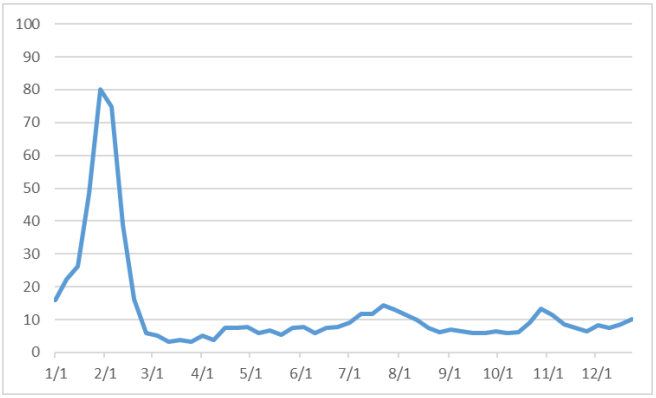

図 2「宮崎　キャンプ」の検索数の推移（5 年間平均）

また，地域特性の抽出の際，その地域における単なる検索数の時系列データを利用すると，その時系列データには一般的な時系列特徴が含まれてしまう．具体例を挙げると，図 2「宮崎　キャンプ」の検索数の推移のグラフには，図 1「キャンプ」の検索数の推移の時系列特徴が含まれている可能性があると考えられる．そこで，本研究ではその地域における単なる検索数の推移の時系列データを用いるのではなく，各地域の時系列データから一般的な時系列データを差し引いた値（差異）を利用する．「キャンプ」の例を用いると，図 3「「宮崎　キャンプ」―「キャンプ」」のようなグラフになる．こうすることで，その事柄においてある地域に限定してみられる地域特性の抽出を行うことができる．

ある地域に限定してみられる時系列特徴の抽出のためには，相関係数を利用した他地域との比較を行う．算出された結果から他の地域との相関が低い，または負の相関をもつ地域では，その事柄に関する地域特性をもつと考えられるため，それらの地域から地域特性となるパターンを抽出した．また，他のキーワードと比較することで，地域特性に関する特徴語の抽出を図った．

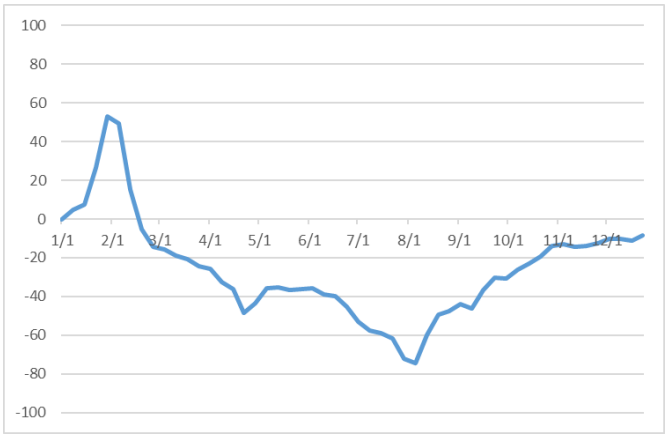

図 3「「宮崎　キャンプ」―「キャンプ」」の差異の推移（5 年間平均）

3．地域特性抽出の手法の提案

3-1．差異を用いた時区間分析

本章では，2 章で挙げた分析手法を用いて，実際に抽出された地域特性について 2 つほど例を挙げて説明したいと思う．

3-1-1．事柄が多面性を持つ例

「キャンプ」の例では，一般的に図 1 のようにゴールデンウィークの時期や夏休みの時期に検索数が増加するグラフがみられる．しかし，宮崎や沖縄といった地域では図 2 のように 2 月に検索数が増加するグラフがみられる．また，図 3 の差異のグラフからも分かるように，全国と比較しても 2 月に差異の値が高くなっている．全国の他の地域と比較しても，全く異なる時期に検索数が増加するという，通常とは異なる時系列特徴をもっているため，宮崎や沖縄といった地域では，「キャンプ」のもつ意味合いが異なる可能性が考えられる．この詳しい分析については，後の 2 節と 3 節で述べる．

3-1-2. 単発的なイベントが発生している例

ここでは,「駅」を例に説明していきたいと思う. 一般的に,「駅」の検索数は図 4 のグラフのように，夏休みや年末年始など, 長期休暇を中心に検索数が増加していることがわかる. しかし，図 5 の「秩父 駅」の例をみてみると春期に検索数の増加がみられる検索数の推移のグラフとなっている．また，図 6 の差異のグラフからも分かるように，全国と比較しても春期に差異の値が部分的に高くなっている．この差異のグラフから,「秩父」地域では「駅」に関する何らかのイベントが春期に行われている可能性が考えられる. 実際に,「秩父 駅」に関係しているイベントとして，秩父鉄道では春期に SL が期間限定で運行されている．このように，時系列特徴の違いから観光情報につながる単発的なイベントも抽出できる可能性があることが理解できる.

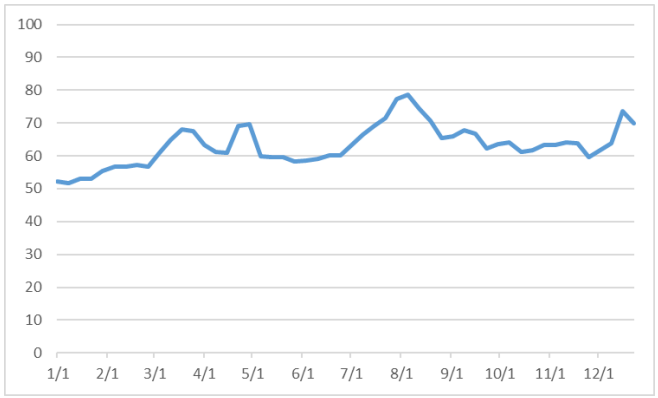

図 4「駅」の検索数の推移（5 年間平均）

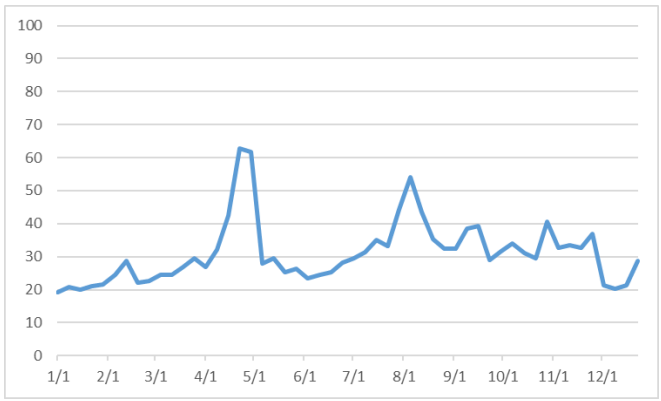

図 5「秩父 駅」の検索数の推移（5 年間平均）

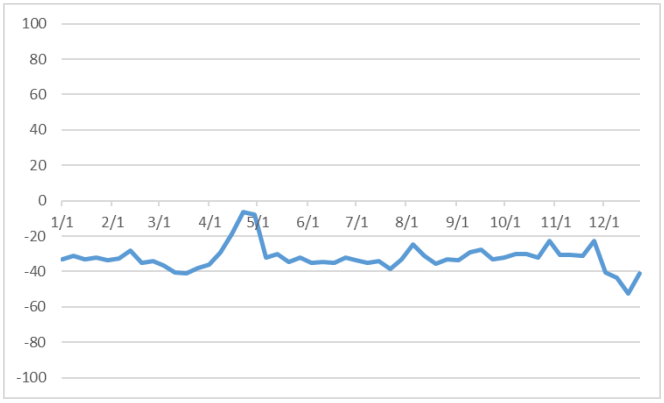

図 6「「秩父 駅」―「駅」」の差異の推移（5 年間平均）

3-2. 相関係数を用いた地域特性の抽出

ある地域に限定された時系列特徴の抽出のために，ある地域の時系列データ「「地域∧事柄」－「事柄」」と他地域の時系列データ「「他地域∧事柄」－「事柄」」の相関係数の値の比較を行う．算出された相関係数を基に，相関がみられなかった，または負の相関をもった地域に関して，その地域に限定される時系列特徴を抽出し，地域特性の抽出を行う.

前節の例で利用した,「宮崎 キャンプ」を例に説明する. 表 1 の相関係数で示されたように,「宮崎 キャンプ」の差異の推移は他地域（栃木・千葉・静岡・岐阜・大分）の差異の推移と比較すると，負の相関または相関があまりみられないことが判断できる．一方,「鹿児島 キャンプ」の差異の推移または「沖縄 キャンプ」の差異の推移のグラフとは正の相関が強い相関係数が算出されている．このことから，全国的な地域とは時系列特徴の違いが認められるが，鹿児島や沖縄と言った九州南部地方では正の強い相関がみられたため，これらの地域に限定してみられる時系列特徴がみられるということが出来る．また，図 3 のグラフから，その地域に限定してみられる時系列特徴が強くみられる時期は，差異の値の絶対値が大きい 2 月や 8 月ということが出来る.

本研究では，この例のように差異の利用と相関係数を用いて他地域と比較することで，その地域に限定される時系列特徴を抽出することができると考えられる．また，抽出の際には，差異の値の絶対値が大きい時期に着目して分析することで，地域特性の抽出を図る.

表 1「キャンプ」の地域間の相関係数

	栃木	千葉	静岡	岐阜	大分	宮崎	鹿児島	沖縄
栃木								
千葉	0.51							
静岡	0.75	0.66						
岐阜	0.65	0.16	0.63					
大分	0.26	-0.18	0.33	0.60				
宮崎	-0.65	-0.41	-0.53	-0.65	-0.18			
鹿児島	-0.22	-0.26	-0.10	-0.11	0.31	0.71		
沖縄	-0.63	-0.41	-0.50	-0.62	-0.14	0.98	0.71	

3-3. 地域特性に関する特徴語の抽出

「宮崎 キャンプ」の例では，「キャンプ」という単語が二面性をもち，宮崎・鹿児島・沖縄といった地域では「キャンプ」という単語が異なる使われ方をしているのではないかということが推測された．本節では，なぜこのような結果が出たのかという裏付けを行うため，同地域におけるランダムに選択された他の事柄との相関を比較することにする．比較した結果，強い正の相関がみられた事柄に関して，関連性の高いキーワードとして抽出を行う．

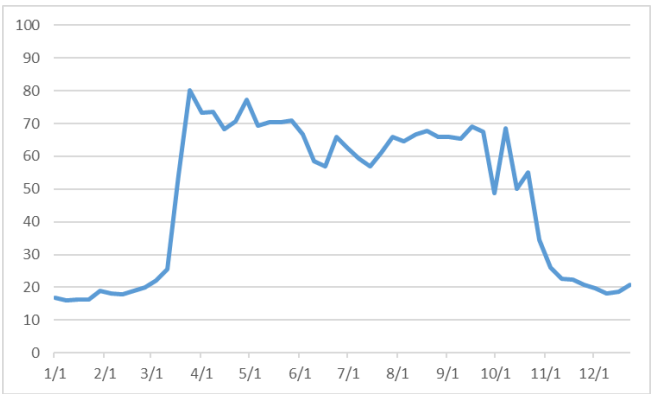

図 7「プロ野球」の検索数の推移（5 年間平均）

「宮崎 キャンプ」の例を挙げると，「宮崎 プロ野球」と強い正の相関がみられた．通常，「プロ野球」は図 7 のように 3 月頃から 10 月頃まで，通して検索数が高い時期がみられる．しかし，宮崎においては図 8 のように 2 月に検索数が増加する．図 9 の差異のグラフをみても理解できるように，「宮崎 プロ野球」は 2 月に地域特性をもつことが考えられる．ここで，図 3「「宮崎 キャンプ」—「キャンプ」」の差異の推移のグラフと図 9「「宮崎 プロ野球」—「プロ野球」」の差異の推移のグラフを比較すると，相関係数は”0.85”という非常に強い正の相関がみられた．

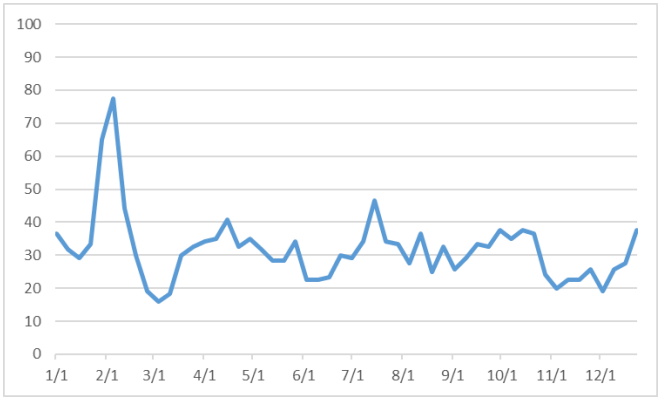

図 8「宮崎 プロ野球」の検索数の推移（5 年間平均）

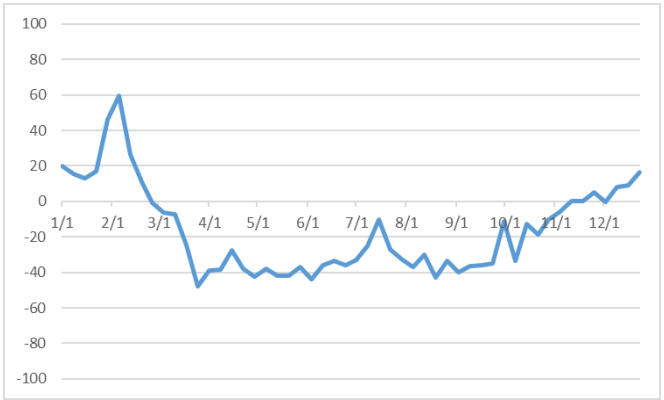

図 9「「宮崎 プロ野球」—「プロ野球」」の差異の推移（5 年間平均）

このように，他の事柄との関係性を調べてみることで，関連性の高いキーワードの抽出が行える．

４．関連研究

インターネット検索サイトでの検索数や，SNS への投稿数などの時空間データを利用して，イベントやスポットなどを抽出する研究は数多く存在する．熊野らの研究[4]では，SNS に投稿された写真に付与されているメタ情報を基に，人気撮影スポットにおける四季に応じた人気のある時期（旬シーズン）の抽出を行った．この際，彼らは旬シーズンの長短や投稿数の大小から推定できるバースト性に着目している．本研究では，地域ごとの検索数を利用することで人気のある時期と事柄の組合せを抽出した．森らの研究[5]では，大量の時空間データに対して相関が高い時系列を抽出してクラスタ化し，ホットスポットとして可視化する手法を提案した．本研究では，ランダムに抽出された事柄の時系列データと比較し，相関が高い事柄を関連性の高いキーワードとして抽出した．

本研究では時空間データを用いた地域特性の抽出を行ったが，時空間データ以外から地域特性を抽出する研究も多く存在する．山岸らの研究[6]では，観光レビューデータからユーザの観光時の行動を確率モデル化し，地域ごとの特性を見出す指標としてスポットランキング手法を提案した．奥らの研究[7]では，IDF 及び地域関連重みから地域限定性スコアを算出し，対象とする空間全体に対して出現頻度が相対的に高い語句を地域限定語句として抽出した．西脇らの研究[8]では，SNS に投稿された位置情報が付与されている写真データを抽出し，クラスタリングを行うことで，その地域における穴場スポットを抽出した．穴場スポット抽出の際，クラスタごとにお気に入り数や写真数を基に穴場スポット度を算出している．白井らの研究[9]では，SNS に投稿された位置情報が付与されている写真データを抽出し，撮影時の方向や時間からホットスポットの抽出と分類，またホットスポット間の関連を調査した．

５．まとめと今後の課題

本稿では，ある地域における特徴的な事柄の検索数が，他の地域におけるその事柄の検索数の推移や周期性などの時系列特徴と違いを持つことを利用した地域特性の抽出を行った．地域特性の抽出の差異には，各地域の時系列特徴から一般的な時系列特徴を差し引いて求めた差異を利用し，別地域との相関係数を算出することで，相関が低い，または負の相関をもつ地域について，その事柄に関する地域特性を抽出した．また，他のキーワードと比較することで，地域特性に関する特徴語の抽出を図った．

今後の課題としては，まず地域の粒度に応じて抽出される内容に違いがあるのかどうかの検討が必要である．また，どういった事柄（キーワード）に関して，地域特性の抽出が行えるのか，どのように地域特性とする事柄（キーワード）を抽出するのかを決定する必要がある．これらを検討・決定したうえで，地域集合・キーワード集合に基づく地域特性の自動抽出を行いたいと考えている．

参　考　文　献

[1]. 日本政府観光局(JNTO) http://www.jnto.go.jp/jpn/statistics/visitor_trends/

[2]. 国土交通省観光庁 http://www.mlit.go.jp/kankocho/siryou/toukei/syouhityousa.html

[3]. Google Trend https://www.google.co.jp/trends/

[4]. 熊野雅仁；岩渕聡；小関基徳；小野景子：木村昌弘「集合知に基づいたポピュラー撮影スポットに関する旬シーズンの可視化」（2013）『芸術科学会論文誌』Vol.13, No.4, pp.218-228

[5]. 森啓太；本田理恵「時空間データからの相関イベントクラスタの共起性の抽出」（2015）DEIM Forum 2015 P2-5

[6]. 山岸祐己；斉藤和巳「観光レビューデータから構築した確率ネットワークによる地域分析」（2016）DEIM Forum 2016 A2-6

[7]. 奥健太；西崎剛司；服部文夫「地域限定性スコアに基づく位置情報付きコンテンツからの地域限定語句の抽出」（2012）『情報処理学会論文誌』データベース, Vol.5, No.3, pp.97-116

[8]. 西脇達也；北山大輔「写真共有サイトを用いた穴場スポットの抽出」（2015）DEIM Forum 2015 P4-5

[9]. 白井元浩；廣田雅春；石川博；横山昌平「ジオタグ付き写真を用いたホットスポットの分類と関連の抽出」（2014）DEIM Forum 2014 E4-3