The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

# ココハドコ?: ユーザの位置が紐付いたウェブ検索履歴の解析

What's Here Like ? Analysis of Web Search Log Based on User's Location

日暮 立 *1　　坪内 孝太 *1
Tatsuru Higurashi　Kouta Tsubouchi

*1ヤフー株式会社
Yahoo Japan Corporation

In recent years, with the spread of GPS-enabled mobile phones, a huge amount of users' historical location data is able to collect. Many studies on modeling with collected location data have been conducted. In this paper, we propose a new method to analyze the area characteristics based on user's web search query logs and location data. As a result of analyzing by the proposed system with various regions and periods, the feature of the area apper in users' search behavior.

## 1. はじめに

近年, GPS 機能を搭載したモバイル端末の普及に伴い, 膨大な量のモバイル端末位置履歴データが蓄積されてきている. モバイル端末に蓄積された位置履歴データを用いて, 都市の分析を行う研究や, ユーザの特徴を解析する研究が盛んである. 都市の分析を行うことで, 都市における需要把握などの都市計画への応用が期待できる. 本研究では位置履歴データを利用し, ユーザのウェブ検索行動から得られる都市や地域の特徴を解析する. Yahoo! JAPAN アプリ利用者の位置履歴データと Yahoo!検索を利用したユーザの検索クエリログを用いて, 地域と期間を指定し機械学習により特徴的な検索ワードを推定するシステムを提案する. 学習の結果得られる, 地域毎の検索クエリの特徴を分析し, 指定期間の違いによる検索クエリの重みの変化から地域内で活動するユーザの興味の変化を分析する.

## 2. 関連研究

位置履歴データを活用して人々の活動パターンを抽出して解析する研究は数多くあり, パターン抽出結果を利用することで都市ごとの特性を把握することができる. [Fan 14] らは 1 日ごとの都市の活動人口のパターンが地域の性質を表現することに着目し, 位置履歴データからその代表的なパターンを抽出した. [Nishi 14] らは階層ベイズによるモデリング手法を用いて活動人口のパターン抽出を行なった. これらのパターン抽出の取り組みを応用して [Shimosaka 15] らは曜日や天気といった要因に対して変化する活動人口の予測に取り組み, [Okawa 17] らは路線毎の交通量予測という問題に取り組んだ. [Yabe 17] らは, 位置履歴データから都市ごとの人々の活動パターンを抽出し, 都市の災害に対する頑強性を示した.

大規模な位置履歴データから都市の状態を把握する研究が盛んであるが, 一方でユーザ個々の位置履歴データに着目しユーザを解析する研究も盛んである. [Wanaka 16] らは位置履歴データからユーザの興味関心を推定する方法を提案した. 位置履歴データからユーザ個人の状態を推定する研究は多いが, 推定の出力結果となるモデルがユーザの検索クエリで構成される推定モデルに関する研究は我々の知る限り存在しない.

連絡先: 日暮 立, ヤフー株式会社, thiguras@yahoo-corp.jp
連絡先: 坪内 孝太, ヤフー株式会社, ktsubouc@yahoo-corp.jp

## 3. 提案システム

本章では, 位置履歴データからユーザの特徴的な検索クエリを予測するためのシステムの説明を行う. 本提案システムでは, 任意の面積の地域と任意の期間を入力とし, 地域の特徴を現す検索クエリを出力とする. 本システムは, 地域と期間を入力し, 位置履歴データからユーザリストを抽出するユーザ抽出部と, 抽出した活動ユーザリストを入力して機械学習によって特徴的な検索クエリを予測するクエリ学習部で構成される. 以下に各システムの詳細を示す.

### 3.1 ユーザ抽出部

ユーザ抽出部は, 地域と期間を入力とし地域内で活動するユーザリストを取得する処理を行う. 以後ユーザ抽出部で出力するユーザを活動ユーザと呼ぶ. ユーザ抽出部では, 任意の面積の地域と任意の期間を入力とし, 蓄積された位置履歴データを元に入力した地域内の活動ユーザと活動ユーザの滞在スコアのリストを取得する. 滞在スコアの計算は位置情報の精度を加味するためにカーネル密度推定法による推定を行う. カーネル密度推定法とは, 密度を計算する地点を中心として任意に指定した検索半径内の点密度を, 計算地点からの距離減衰効果による重み付けを伴って計算する手法である.

あるユーザ蓄積されている位置履歴データのタイムスタンプを $T = \{t_1, t_2, ..., t_n\}$ とし, 時刻 $t_i$ の緯度を $x_{t_i}$, 経度を $y_{t_i}$, 位置情報の精度を $a_{t_i}$ とし, ユーザの位置履歴データの集合を $D = \{(x_{t_i}, y_{t_i}, a_{t_i}) \mid t_i \in T\}$ とする. $(x_{t_i}, y_{t_i})$ が指定地域に含まれる場合, ユーザの滞在スコアは 1 となる. $(x_{t_i}, y_{t_i})$ が指定地域に含まれない場合, 指定地域の中心点と $(x_{t_i}, y_{t_i})$ の距離を $d_{t_i}$ とすると, ユーザの滞在スコアはカーネル密度推定法により,

$$\begin{cases} \frac{1}{T}\sum_{i=1}^{T} K[\frac{d_{t_i}}{a_{t_i}}] & (a_{t_i} \geq h) \\ \frac{1}{T}\sum_{i=1}^{T} K[\frac{d_{t_i}}{h}] & (otherwize) \end{cases} \quad (1)$$

となる. $d_{t_i}$ の計算はヒュベニの距離計算式を用い, カーネル密度推定法のカーネル関数 $K$ はガウス関数を採用した.$h$ はバンド幅である. 蓄積されているユーザの位置履歴データから点密度を計算し, 点密度が 0 より大きい上位の $N$ 件のユーザを活動ユーザとして抽出する.

### 3.2 クエリ学習部

クエリ学習部では, ユーザ抽出部で抽出した活動ユーザリストとユーザ抽出部で指定した期間を入力として, 活動ユーザら

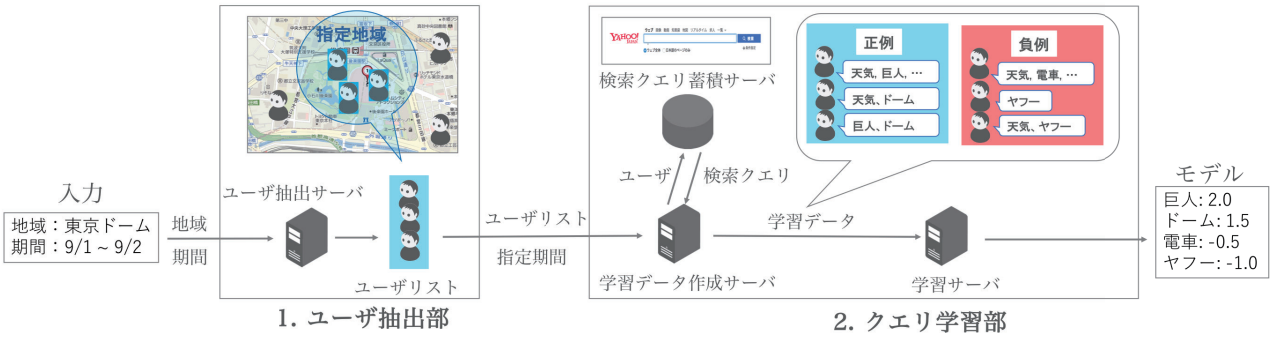


図 1: 提案システム概要

の特徴的な検索クエリを推定する．ユーザ抽出部にて出力した $N$ 件の活動ユーザリストを $U_{pos} = \{p_1, p_2, ...p_N\}$ とする．また, $U_{pos}$ に含まれない, 指定期間に Yahoo!検索を利用したユーザをランダムに $N$ 件サンプリングし、$U_{neg} = \{n_1, n_2, ...n_N\}$ とする．

ユーザ $p_i$ と $n_i$ が指定期間に Yahoo!検索にて検索を行なった検索クエリのリストを検索クエリログ蓄積サーバから取得し, 取得した検索クエリの総数を $Q$ とする．各ユーザ $p_i$ と $n_i$ が, 指定期間に検索したクエリは 1, それ以外のクエリを 0 とする $Q$ 次元 の one-hot ベクトル $v \in \{0,1\}^Q$ を特徴量ベクトルとする．ユーザ抽出部によって抽出した活動ユーザリスト $U_{pos}$ を元に生成した特徴量ベクトルの集合 $D_{pos} = \{v_{p_i} \mid p_i \in U_{pos}\}$ を学習データの正例とし, ランダムにサンプリングしたユーザリスト $U_{neg}$ を元に生成した特徴量ベクトル $v$ の集合 $D_{neg} = \{v_{n_i} \mid n_i \in U_{neg}\}$ を学習データの負例とする．

生成した正例 $D_{pos}$ と負例 $D_{neg}$ を学習データとし,LIBLINEAR[*1] を使用した二値分類学習を行う．学習の結果の学習モデルから得られる,$Q$ 個の素性として入力された検索クエリとその重みを出力する．学習の正規化パラメータ $C$ は交差確認でパラメータ探索を行い, 精度が最も良くなる $C$ を採用するようにしている．

重みが高い素性は指定した地域内の活動ユーザ $U_{pos}$ が特徴的に検索している検索クエリであることがわかり, 重みが低い素性はユーザ $U_{neg}$ が特徴的に検索している検索クエリであることがわかる．

## 4. データ

本章では, 提案システムで使用するデータに関する詳細を示す．

ユーザ抽出部にて利用した位置履歴データは， Yahoo! JAPAN アプリ [*2] を Andorid また iOS のスマートフォンにインストールし, 位置情報を利用したサービス提供を受けるために「位置情報に基づいたコンテンツを表示する」という項目をオンにした Yahoo! JAPAN ID ログインユーザのデータを利用した．データはタイムスタンプと緯度経度と位置情報の精度の情報を持ち, 日本全国のユーザから取得されている．データは基本的に移動している状態の端末や基地局が切り替わった端末から取得されるため, 人々の動きを表したデータと言える．スマートフォンで取得した位置履歴データは個人の位置の正確な特定を避けるために 100m メッシュデータに変換している．利用時には k-匿名性を担保するため提案システムの抽出部において抽出人数には制限をかけている．

Android 端末のスマートフォンと iOS 端末のスマートフォンは位置情報の取得頻度が異なる．提案システムで利用する位置履歴データのタイムスタンプの間隔を共通化するために位置履歴データの内挿を行う．あるユーザの位置履歴データのタイムスタンプを $t$, 緯度を $x$, 経度を $y$, 精度を $a$ とし, 連続して蓄積された $t_0 < t_1$ となる履歴データをそれぞれ $(t_0, x_0, y_0, a_0)$, $(t_1, x_1, y_1, a_1)$ とする． $S$ を 0 より大きい任意の秒数として $t_1 - t_0 > S$ の時, 線形補間を行う．$S$ はシステム内で定義されるハイパーパラメータである．$\{i \in \mathbb{N} \mid i < (t_1 - t_0)/S\}$ に対して, タイムスタンプが $t_i = t_0 + S * i$ 時の内挿データの $x_i$,$y_i$, $a_i$ は以下のように算出される．

$$x_i = x_0 + (x_1 - x_0) * \frac{t_i - t_0}{t_1 - t_0} \tag{2}$$

$$y_i = y_0 + (y_1 - y_0) * \frac{t_i - t_0}{t_1 - t_0} \tag{3}$$

$$a_i = a_0 + (a_1 - a_0) * \frac{t_i - t_0}{t_1 - t_0} \tag{4}$$

クエリ学習部にて利用しているユーザの検索データは, ウェブブラウザもしくは携帯端末にインストールされているアプリで利用された Yahoo! JAPAN ID ログインユーザの Yahoo! 検索 [*3] のログを利用した．

## 5. 実験と結果

実験では $S = 300$ として内挿を行なった位置履歴データを利用した．提案システムのユーザ抽出部において, カーネル密度推定法のバンド幅 $h = 250$, 抽出数 $N = 3000$ として活動ユーザを抽出し, クエリ学習部に使用した．3 つのケースについて, 提案システムによる解析例を以下に示す．

### 5.1 北海道地震 (北海道厚真町)

表 1 の例は, 北海道勇払郡厚真町を含む 1600km$^2$ の地域を指定し, 2018 年 9 月 6 日の 4 時から 2018 年 9 月 7 日の 17 時の期間を指定して提案システムによる学習を行った時の重みが上位の素性と重みが下位の素性の一部である．2018 年 9 月 6 日の 3 時 7 分には, 北海道胆振地方中東部を震源として地震が発生しており, 厚真町では最大震度 7 が観測されている．

*1 https://www.csie.ntu.edu.tw/ cjlin/liblinear/
*2 https://promo-mobile.yahoo.co.jp/yjapp/
*3 https://search.yahoo.co.jp/

表 1: 厚真町に滞在していたのユーザのクエリ

| 重みが高い検索クエリ | | 重みが低い検索クエリ | |
|---|---|---|---|
| 検索クエリ | 重み | 検索クエリ | 重み |
| 新千歳空港 | 1.529 | 地下鉄 | -2.57 |
| 道路交通情報 | 1.426 | 意味 | -2.118 |
| ラジコ | 1.408 | 旭川 | -1.779 |
| 運行状況 | 1.353 | iphone | -1.397 |
| ガソリンスタンド | 1.238 | ユーチューブ | -1.386 |
| 北海道電力 | 1.232 | 吉澤ひとみ | -1.343 |
| 苫小牧市停電情報 | 1.198 | au | -1.249 |
| 鍋でご飯を炊く方法 | 0.97 | おしゃれ | -1.242 |

学習結果から, 指定した期間に北海道厚真町周辺に住む住民は「道路交通情報」,「ガソリンスタンド」や「北海道電力」などの検索クエリの重みが強く, 災害時にはこれらの情報を特に必要としていることがわかり, 一方で「iphone」「ユーチューブ」「おしゃれ」などの日常の娯楽を連想させる検索クエリは重みが低く, このような情報は災害時には特に必要としていないということがわかる. 実験から, 提案システムによる検索クエリを予測することで, 指定地域の災害状況を把握し住民が被災時にどのような状態かを素早く判断することができる.

### 5.2 安室奈美恵ライブ (沖縄県)

表 2 は沖縄県本島全域を含むを 8400km$^2$ を地域として指定し, 指定期間を変化させた時の提案システムによる学習を行った時の重みが上位の素性の一部である. 指定期間はそれぞれ 2018 年 9 月 12 日,2018 年 9 月 15 日,2018 年 9 月 18 日から 1 日間である. 沖縄県では歌手の安室奈美恵が引退ライブを 2018 年 9 月 15 日に行なっており, 表 2 ではすべての日に「安室奈美恵」という素性が重みが高く, ライブに関する特徴がすべての日に現れている. 各日付の「安室奈美恵」という検索クエリの重みに注目すると, ライブの前からライブ当日に向けて「安室奈美恵」という素性の重みは高く現れているが, ライブ終了後は重みが低くなっている. 一方で「沖縄」のような地名を表す重みは指定期間の影響を受けることなく, 常に高い重みで現れている. 提案システムの出力結果からはイベントによる瞬間的に現れる検索クエリと, 常時現れている検索クエリが存在することがわかる. 瞬間的に現れる検索クエリを分析することで指定地域内のユーザの興味の変遷の把握が期待でき, 常時現れる検索クエリを分析することで指定地域の特性の把握が期待できる.

### 5.3 東京ドームのイベント (東京都)

表 3, 表 4, 表 5 は東京ドームを含む地域 0.1km$^2$ を指定し, 指定期間を変化させた時の提案システムによる学習を行った時の重みが上位の素性の一部である.

指定期間はそれぞれ, 2018 年 11 月 17 日から 2019 年 1 月 20 日の間の土日の 2 日間である. 東京ドームでは週末には様々イベントが開催されており, アニメ「ラブライブ!サンシャイン!!」の音楽イベントやアイドルグループ「嵐」「Kis-My-Ft2」「SUPER JUNIOR」のドームライブ [*4], また音楽イベント以外にも全国の祭りとご当地の食を堪能できる「ふるさと祭り 2019」[*5]、野球のファンイベント [*6]、プロレスイベント [*7] が行われていた. 提案システムによって出力される検索クエリを考察すると, 東京ドーム周辺の地名以外に開催されているイベントに関連する検索クエリが特徴として現れている. 提案システムによって出力される検索クエリから, 指定した地域で行われているイベントを推定することが可能であることがわかる. また, 表 4 のふるさと祭り東京の結果では「座席表」「時刻表」という検索クエリが特徴として強く現れている. ふるさと祭り東京は全国の祭りと食を堪能できるイベントとなっているため, 全国から来場者が訪れる. 来場者が全国各地から東京ドームに訪れる際に, 新幹線や特急列車を利用している可能性が高いことが結果から考がえられる. このように, 本提案システムからは, イベント抽出以外にも移動経路や移動方法の分析も期待できる.

∗4 https://www.tokyo-dome.co.jp/dome/event/artist/
∗5 https://www.tokyo-dome.co.jp/furusato/
∗6 http://meikyu-kai.org/bf2018/
∗7 https://www.njpw.co.jp/tornament/163210

## 6. まとめ

本論文では, Yahoo! JAPAN アプリを利用しているユーザの携帯端末の位置履歴データを利用して指定地域の活動ユーザを抽出し, ユーザの検索クエリを特徴量としてロジスティック回帰により学習することで, 指定地域の特徴的なウェブ検索クエリを学習するシステムを提案した. 様々な指定地域に対して提案システムによる分析を行なった結果, 指定した地域にて行われたイベントに関する検索行動が検索クエリに現れることが示された. またイベント以外にも, 指定地域で活動ユーザが必要としている情報を検索クエリから抽出することが可能であることがわかった. 同じ指定地域において指定期間を変化させたところ, 提案システムによって抽出された検索クエリの重みを比較することで, 指定地域の活動ユーザの興味の変化を推定することが可能であることが示された.

地域ごとのウェブ検索クエリの解析の結果は, リスティング広告などのインターネット広告の配信への応用が期待される. 今後は提案システムによって出力された地域毎のウェブ検索クエリモデルの実活用とその効果の検証を検討したい.

## 参考文献

[Fan 14] Fan, Z., Song, X., and Shibasaki, R.: CitySpectrum: A non-negative tensor factorization approach, UbiComp (2014)

[Nishi 14] Nishi, K., Tsubouchi, K., and Shimosaka, M.: Extracting Land-use Patterns Using Location Data from Smartphones, URB-IOT (2014)

[Okawa 17] Okawa, M., Kim, H., and Toda, H.: Online Traffic Flow Prediction Using Convolved Bilinear Poisson Regression, MDM (2017)

[Shimosaka 15] Shimosaka, M., Maeda, K., Tsukiji, T., and Tsubouchi, K.: Forecasting Urban Dynamics with Mobility Logs by Bilinear Poisson Regression, UbiComp (2015)

[Wanaka 16] Wanaka, S. and Tsubouchi, K.: Location History Knows What You Like: Estimation of User Preference from Daily Location Movement, Urb-IoT (2016)

[Yabe 17] Yabe, T., Tsubouchi, K., and Sekimoto, Y.: CityFlowFragility: Measuring the Fragility of People Flow in Cities to Disasters Using GPS Data Collected from Smartphones, Ubicomp (2017)

表 2: 沖縄に訪れたユーザの検索クエリ

| 安室奈美恵のライブ 3 日前 | | 安室奈美恵のライブ当日 | | 安室奈美恵のライブ 3 日後 | |
|---|---|---|---|---|---|
| 検索クエリ | 重み | 検索クエリ | 重み | 検索クエリ | 重み |
| 沖縄 | 3.108 | 沖縄 | 3.644 | 沖縄 | 2.711 |
| **安室奈美恵** | **1.411** | 那覇 | 2.555 | 沖縄県 | 0.849 |
| 那覇 | 1.17 | 那覇市 | 1.772 | 古賀淳也 | 0.763 |
| 那覇市 | 0.763 | **安室奈美恵** | **1.769** | 那覇市 | 0.445 |
| 沖縄県 | 0.672 | ほっともっと | 1.757 | 那覇 | 0.417 |
| 直接謝罪 | 0.67 | 宜野湾 | 1.596 | 沖縄市 | 0.157 |
| 台風情報 | 0.404 | 地震速報 | 1.388 | 山本 kid 徳郁 | 0.116 |
| 沖縄宝島 | 0.351 | 読谷 | 1.38 | 沖縄タイムス | 0.116 |
| 巨人杉内 | 0.335 | 今週の急上昇ワード | 1.344 | **安室奈美恵** | **0.041** |
| ひびきわたる | 0.202 | 浦添 | 1.189 | 琉球新報 | 0.035 |

表 3: 東京ドームを訪れたユーザの検索クエリ (2018 年 11 月付近)

| ラブライブ音楽イベント | 野球のファンイベント | SUPER JUNIOR のライブ |
|---|---|---|
| **ラブライブサンシャイン** | **名球会フェスティバル 2018** | 新宿 |
| 東京ドーム | 東京 | **super** |
| **ラブライブ** | 東京ドーム | 東京ドーム |
| **aqours** | 池袋 | 東京 |
| 新宿 | jra | 水道橋 |
| アクセス | アクセス | 池袋 |
| 東京 | 銀座 | 死去 |
| 文京区 | 上野 | 今週の急上昇ワード |
| 秋葉原 | 死去 | 渋谷 |
| 紅白 | **名球会** | 秋葉原 |

表 4: 東京ドームを訪れたユーザの検索クエリ (2018 年 12 月付近)

| 嵐のライブ | Kis-My-Ft2 のライブ | 嵐のライブ |
|---|---|---|
| **嵐** | **キスマイ** | 東京ドーム |
| 東京ドーム | 東京ドーム | **嵐** |
| 銀座 | アクセス | 銀座 |
| 今週の急上昇ワード | 東京 | 神楽坂 |
| 池袋 | 秋葉原 | 水道橋 |
| 水道橋 | 東京駅 | 東京 |
| ヌード | 池袋 | みずほ銀行 |
| 東京 | 大恋愛 | 浅草 |
| 浅草 | 新宿 | 今週の急上昇ワード |
| 東京駅 | 千葉 | 鶯谷 |

表 5: 東京ドームを訪れたユーザの検索クエリ (2019 年 1 月付近)

| 新日本プロレス興行 | ふるさと祭り | ふるさと祭り |
|---|---|---|
| 東京ドーム | **ふるさと祭り東京 2019** | **ふるさと祭り東京 2019** |
| 池袋 | 東京ドーム | 銀座 |
| news | 新宿 | 新宿 |
| 祝日 | 今週の急上昇ワード | 時刻表 |
| 東京 | アクセス | 悲痛コメント |
| **新日本プロレス** | 池袋 | 渋谷 |
| 東京ドームシティ | 天地総子 | ランチ |
| 高校サッカー | 東京 | 座席表 |
| 箱根駅伝 | 北千住 | 東京ドーム |
| 上野 | 東京駅 | 浅草 |