情報処理学会論文誌　コンシューマ・デバイス & システム　Vol.4　No.1　1–11 (Apr. 2014)

コンシューマ・サービス論文

ソーシャル観光マップ ——ソーシャルデータからの観光スポット抽出

荒川 豊[1,a)]　タチアーナ シェフラー[2]　ステファン バウマン[3]　アンドレアス デンゲル[3]

受付日 2013年9月18日, 採録日 2014年1月25日

概要：本論文では，位置情報付きのソーシャルデータ分析に基づくソーシャル観光マップの構築に向け，都市の人気スポットをその正確な名前とともに抽出する仕組みを提案する．人気スポットの名前を推定する手法として，Foursquare などの複数のチェックインサービスから得られる情報を用いることで，従来のタグ分析手法と比較して，正確な表記の名前を得ることを可能とする．本手法を用いることにより，小さなデータセットを用いた場合であっても，正確性の高い名前付けが可能となり，分析速度の高速化と両立できることを明らかにする．実際に，5 都市 436 万枚の写真データを Flickr から収集し，従来方式（枚数による順位付け＋タグ分析による意味付け）と提案方式（枚数と時間分散による順付け＋チェックインサービスを用いた意味付け）による観光スポット上位 10 件を選出，比較することで．提案方式が小さなデータセットであっても正確な名前を持つ観光スポット一覧を生成可能であることを明らかにする．

キーワード：ソーシャルデータ分析，観光，地図，Flickr，Foursquare，シティーマップ

Social Tourist Map —Discovering Popular Point of Interests for Tourism from Social Data

Yutaka Arakawa[1,a)]　Tatjana Scheffler[2]　Stephan Baumann[3]　Andreas Dengel[3]

Received: September 18, 2013, Accepted: January 25, 2014

Abstract: This paper proposes a method to retrieve attractive sightseeing spots of cities through a social data analysis. Especially focusing on how to assign an appropriate name for each clustering result. Our method that combines several Place APIs can estimate more proper name than a conventional method based of a tag analysis, even if the size of dataset size is small. Furthermore, the calculation speed of our proposed method is faster than those of tag analysis. By using our collected data, more than 4 million geo-tagged photos of 5 cities from Flickr, we show our proposed method can semi-automatically generate sightseeing map with appropriate spot names.

Keywords: social data analysis, sightseeing, map, Flickr, Foursquare, city map

1. はじめに

近年，GPS を搭載したスマートフォンが広く普及し，位置情報サービスがこれまで以上に活発になっている．位置情報サービスとは，位置に応じて適切な情報を提供するサービスであり，その場の地図を表示するだけでなく，近隣のレストラン推薦や位置に応じた辞書切替え [1] など位置を利用して関連情報を推薦するさまざまなものを含む．位置に応じた情報を提供するためには，推薦候補となる情報にその位置情報を含む必要があり，近年ではレストラン情報であっても，店名や電話番号，営業時間と並んで，その緯度・経度情報がデータベースに登録されるようになっている．写真に関しては，写真 SNS（Social Network

[1] 奈良先端科学技術大学院大学
Nara Institute of Science and Technology, Ikoma, Nara 630–0192, Japan
[2] University of Potsdam, 14476 Potsdam, Germany
[3] DFKI GmbH, 67663 Kaiserslautern, Germany
[a)] ara@is.naist.jp

Service）である Flickr [*1] が 2006 年から位置情報の登録機能を提供していることから，スマートフォンが普及する前の写真でも位置情報を含む写真が多く蓄積されている．近年では，Twitter [*2] や Facebook [*3] においても，投稿に対してジオタグと呼ばれる位置情報のタグを付与することが可能となっており，一般的なユーザが生成する日々のデータも位置情報を含むようになってきている．その最たる例は，「チェックイン」と呼ばれる SNS によって生成されるデータである．Foursquare [*4] によって広く普及したチェックインは，利用者が訪問地に足あとを残し，その履歴によってバッジなどのインセンティブを受け取るものであるが，現在は Facebook もチェックイン機能を提供しており，日々，位置情報の足あとが蓄積されている．

そして，このような位置情報を含むソーシャルデータ（以降，ソーシャルジオデータ）を分析することによって，従来のセンサネットワークでは発見できなかった実世界情報を抽出する研究が注目されている．たとえば，2010 年にニフティが公開した「みんなの花粉症なう！ β」は，Twitter のジオタグとその文面から花粉症の到来を可視化したサービスである．他にも，実世界でのイベント情報を抽出する研究 [2], [3] なども行われている．また，2011 年の東日本大震災時に開発された sinsai.info [*5] は，Twitter 上の情報を含むさまざまな情報を地図上に可視化したものであり，ソーシャルジオデータを有効に利用した一例である．

こうしたソーシャルジオデータ分析の中で，最も研究が進んでいる分野の 1 つが「観光」である．ソーシャルジオデータを分析して観光地図を生成する研究は，Chen らが 2009 年に取り組んでいる [4]．この研究は，Flickr 上の位置情報付き写真をデータとして，(1) クラスタリングによる POI（Point of Interest）の抽出，(2) 抽出された POI に対する画像分析による意味付け，(3) クラスタの代表画像選出，(4) 人気度を考慮した地図上への画像配置，から構成されている．同時期に，Crandall ら [5] は，全世界を対象として収集した膨大な数の Flickr 画像から，地球上で最も写真が撮られるエリアやそのエリアにおける人気スポットの抽出を行っている．その後，これらの研究の発展として，都市内におけるクラスタ間の遷移をマルコフモデルで解析し，フォトグラファの興味と時間制約を満たしたルートを推薦する手法 [6] や，ユーザのルートのランキング手法 [7]，Foursquare の足あとから観光ルートを分析する手法 [8] など，種々の研究がなされている．

本論文は，このようなソーシャルジオデータ分析に基づいた観光情報の抽出に関する研究であり，特に Chen らの研究 [4] における，(2) POI に対する意味付けと，(4) 人気度を考慮した地図上への画像配置，に焦点を当てる．(1) に関しては従来手法を用い，(3) に関しては今回は取り扱わない．(4) に関しても画像配置そのものではなく，配置する際に考慮する観光スポットの人気度をいかに定量化するかという点に焦点を絞る．さらに，分析の高速化を測るため，このような情報を抽出するために要するデータセットのサイズについて検証する．

POI に対する意味付け手法は，ソーシャルジオデータに対してクラスタリングを適用した場合に得られる結果（クラスタの中心座標）に対して，有意な名前を付与するものであり，今回は観光情報に的を絞っているが，前述した実世界イベントの抽出や災害情報の抽出に対しても応用可能な技術である．また，本論文で提案する意味付け手法を用いることにより，データセットサイズを削減した場合も，名前の正確性を損なうことなく，計算時間を短縮できる点も共通の利点となる．

以降，2 章において従来の意味付け手法とその課題をまとめ，3 章において本論文で利用するさまざまな関連技術について説明する．そして 4 章で事前実験について言及し，5 章で提案手法について説明する．6 章でいくつかの検証例を示したうえで，7 章で総括する．

2. 従来手法とその課題

POI に対する意味付けは，写真そのものを分析しその被写体から推定する手法（Visual Information）[4] と，写真に付与されたタグ情報から推定する手法（Textual Information）[6]，その両者を組み合わせた手法 [5] がこれまで提案されている．Visual Information を用いた手法は，SIFT 特徴 [9] を用いて，当該写真と前もって構築された画像データベース内の画像群との類似度を評価するものである．画像分析による POI の推定は，教師データの選定やその学習コストが膨大であるうえ，教師データに含まれない未知の POI の推定は行えないという点から，本論文では Textual Information のみを利用することを前提とする．

分析対象データの取得元である Flickr における代表的な Textual Information は，各写真に付与された「タグ」である．クラスタ内に含まれるすべての写真に含まれるすべてのタグの中から，そのクラスタを代表するタグを選出する手法として，以下の式により求められるタグスコア $T(V)$ を用いた手法 [5], [6] が提案されている．

$$T(V) = P(m \mid V) = \frac{N(V,m)}{N(V)} \tag{1}$$

ここで $N(V,m)$ は，クラスタ m においてタグ V を含む写真の枚数であり，$N(V)$ はすべての写真の中でタグ V を含む写真の枚数である．

この式では，クラスタ m に多く含まれるタグのうち，全体のクラスタにも多く含まれるタグのスコアが小さくなり，タグスコアが大きいタグほどクラスタ m にだけよく現れ

[*1] http://www.flickr.com/
[*2] http://twitter.com/
[*3] http://www.facebook.com/
[*4] http://foursquare.com/
[*5] http://www.sinsai.info/

るタグとなる．しかしながら，Flickr の各写真に付与されたタグは，不正確なものが多く，得られた代表的なタグをそのまま観光スポット名として利用することは難しい．また，計算速度の観点から，分析対象となるデータセットサイズは小さいほうが好ましいが，データセットサイズを小さくした場合，珍しいタグ（ノイズ）が代表的なタグとして選出される可能性が増大するという問題もある．

一方，画像サイズに反映する POI の人気度に関しては，Chen ら [4] は，人気度に応じて，画像サイズを変化させることを提案しているが，人気度の定量化には言及しない．また，Crandall ら [5] のランキングは，すべて写真の枚数に基づいており，写真が多く撮られた場所を人気エリア（スポット）と定義しているが，観光とは関係のないイベント（特に近年のソーシャルネットワークでの情報伝播を狙ったイベントや，フォトエキスポといった大規模な展示会など）などで写真がたまたま多く撮影されることもあり，「観光」を主眼に考えた場合，必ずしも枚数がそのスポットの人気度を表しているとは限らない．

3. 関連技術

ここでは本論文に関連する研究として，クラスタリング手法である Mean Shift 法と，チェックイン候補を取得するためのリバースジオコーディングとその API（Application Programming Interface）について説明する．

3.1 Mean Shift 法

Mean shift 法 [10] は，主に画像分析 [11] や物体追跡に用いられてきたクラスタリング手法であるが，Crandall らが文献 [5] において，緯度・経度からなる空間情報に対しても適用可能であることを示してからは，いくつかの研究 [6], [7] で空間情報のクラスタリングに用いられている．Chen ら [4] が用いている，k-means 法と比較して，Mean Shift 法はクラスタ数 k を事前に決定する必要がないというメリットがある．その他の空間情報のクラスタリング手法としては，Kisilevich らによる p-DBSCAN [12] や，Yang らによる Self-tuning Spectral Clustering [13] などが提案されているが，これらの手法は，POI の大きさや形状の違いを考慮したクラスタリング手法である．そのため本論文では，パラメータが少ない点を重視し，Mean Shift 法を適用する．

Mean Shift 法では，Bandwidth w と呼ばれる 1 つのパラメータのみを設定し，ある観測点の点 x から半径 w に含まれる点の重心（平均値）を次の観測点として，密度分布関数の極大値を検出する．観測点 x における Mean Shift ベクトルを m は下記のように定義できる．

$$m_{h,G}(x) = \frac{\sum_{i=1}^{n} x_i g\| (x - x_i)/w \|^2}{\sum_{i=1}^{n} g\| (x - x_i)/w \|^2} - x \qquad (2)$$

この式において，x_i は半径 w に含まれる観測点を示し，g は G で指定されたカーネル関数を表す．カーネル関数としては，一様カーネル [5] やガウシアンカーネル [6] が用いられており，本研究では後者のガウシアンカーネルを採用する．これは，観光スポットの中心部ほど写真が多いという仮定に基づいている．

Mean Shift 法は，任意の観測点 $x(1)$ から計算を始め，下記の式に基づいて観測点を移動しながら，Mean Shift ベクトルが 0 に収束するまで計算を繰り返す．

$$x_{(i+1)} = x_i + m_{h,G}(x_i) \qquad (3)$$

空間情報分析においては，Bandwidth $w = 0.001$ は約 100 m，$w = 1$ は約 100 km を表す．Crandall ら [5] は，全世界から都市を抽出する場合に $w = 1$，各都市のスポットを抽出する場合に $w = 0.001$ を用いている．また，スポットの抽出を目的とした Yang らの研究 [13] では，$w = 0.001$ としている．一方，ルート分析と推薦に関する Kurashima らの研究 [6] では，$w = 0.0001$ ときわめて小さな値（約 10 m 相当）を用いている．本研究は，スポットの抽出を目的としていることから，以降の章では $w = 0.001$ を用いる．

3.2 リバースジオコーディング API について

位置情報サービスの普及にともない，文字列として住所を地図上に投影可能な座標（緯度・経度）情報に変換する，ジオコーディング（Geocoding）と呼ばれるサービスが普及してきている．同時に，座標情報から，住所，あるいはスポット名や店名といった人間が認識可能な文字列情報に変換する，リバースジオコーディング（Reverse Geocoding）というサービスも普及している．これらのサービスは，一般的に，Web API を介して提供されており，一般ユーザからも利用することが可能となっている．特に，“GeoNames [*6]” と “OpenStreetMap [*7]” は有名な公開サービスであり，巨大な位置情報データベースが無償で公開されている．

また，近年では「チェックイン」という，その場所に来たことを SNS 上で知らせるサービスが広く普及している．これは，Foursquare が 2009 年に始めたサービスであるが，現在では Google や Facebook といったメジャーな企業が同様のサービスを提供している．この「チェックイン」サービスでは，ユーザに対して，その位置におけるチェックイン対象となる候補を一覧表示する．その際に用いられるのが前述したリバースジオコーディング機能であり，ユーザの所望するチェックイン候補をより上位に提示した方が利便性が向上することから，各社は，位置情報データベースとそこからの選出アルゴリズムを競い合っている．**表 1** は，有名なリバースジオコーディング API を比較したものである．これ以外にも，De Choudhury ら [14] が用いてい

[*6] http://www.geonames.org/
[*7] http://www.openstreetmap.org/

る Yahoo GeoPlanet API[*8]や，レストラン情報なども網羅した Yelp API[*9]，OpenStreetMaps のデータを利用した CloudMade API[*10]など，さまざまな API が存在するが，サービスの持続性[*11]などをふまえ，本論文では表に示す 3 つの API を利用する．

まず, Foursquare は, 初期データとして前述の GeoNames のデータを用いているが，ユーザが新しい POI を自由に登録できるという特徴がある．チェックインの種類や回数に応じてバッジと呼ばれるインセンティブを付与したり，ある POI に対して最も頻繁にチェックインするユーザにメイヤーと呼ばれる称号を与えたり，新たな POI の追加に対してポイントを付与したりと，ゲーミフィケーションによって，位置情報データベースに登録されていない未知の POI がユーザによって次々と追加される仕組みになっており，登録されているデータ数は最大である．2013 年 3 月のニュース[*12]によると 5,000 万件以上の POI が登録されている．しかしながら，ユーザによって登録される情報は，その粒度や表記も統一されておらず，登録されているデータ数が多い方が必ずしも良いとは限らない．鉄道駅を例にとると，ある駅では各乗車ホームが別の POI として登録されていたり，複数の路線が乗り入れる駅では駅名に路線名まで含んでいたりすることも多い．また，海外の例では，フランクフルト空港（Frankfurt International Airport）という 1 つの POI に対して，“Frankfurt Airport”，“Frankfurt Flughafen”，“Flughafen Frankfurt am Main” とさまざまな言語が混在していることもある．一方，Facebook が提供する Graph API は，Factual[*13]の商用データベースを利用している．Facebook は，ユーザによる POI の新規登録を許可していないため，Foursquare と比較して，登録されているデータ数は少ないものの，正確性の高い情報のみが登録されている印象である．ちなみに，Google が提供する Places API の基盤データは不明であるが，Google Maps の資産を活用していると考えられる．

もう 1 つの大きな相違点は，カテゴリ指定が可能か否かである．膨大な POI データベースから適切な情報を抽出したい場合，目的やアプリケーションに応じてカテゴリを限定することによって精度を改善できると期待できる．今回取り上げた 3 つの API の中で，このカテゴリを指定可能な API は Foursquare API と Google Places API であるが，両者のカテゴリ分類は大きく異なるという問題がある．具体的には，Foursquare のカテゴリは，9 つの主カテゴリと，その下に含まれる多数のサブカテゴリから構成される階層的なカテゴリとなっており，主カテゴリを指定することによって，下位のサブカテゴリすべてを指定することが可能となっている．一方，Google はフラットな 126 のカテゴリから構成されている．

各 API の共通点としては，各社独自のアルゴリズムに基づいた重要度（人気度）に基づいて出力順位が決定されるという点である．これは，スマートフォンで得られる位置情報の精度がそれほど高くないことから，緯度経度から得られる距離が近いからといって，必ずしも実際に距離が近いとは限らないためである．しかしながら具体的なアルゴリズムはすべて不明である．なお，Google Places API に関しては，距離に基づいたアルゴリズムを指定して出力を得ることも可能である．

表 1　リバースジオコーディング API の比較
Table 1　Comparison of Reverse geocoding APIs.

名前	POI の登録	カテゴリ指定	出力アルゴリズム
Foursquare API	Yes	Yes	Popularity
Facebook Graph API	No	No	Popularity
Google Places API	No	Yes	Prominence or Distance

*8 http://developer.yahoo.com/geo/geoplanet/
*9 http://www.yelp.com/developers/
*10 http://cloudmade.com/
*11 巨大なデータを維持と継続的な情報更新には膨大な費用がかかるため，いつの間にかサービスを停止あるいは会社が消滅している場合が多い．
*12 http://www.blogherald.com/2013/03/11/foursquare-possibly-switching-focus-from-check-in-to-api-data/
*13 http://www.factual.com/

4. 事前実験

4.1 カテゴリ設定に関する事前実験

「チェックイン」を行うためにはインターネットへのアクセスが必要であるため，無料 WiFi を提供するマクドナルドやスターバックスなどが，チェックイン対象の上位に抽出されることがある．カテゴリを指定することによって，このような目的（今回は観光）に関係のない情報を低減させることができると考えている．今回は，著者の主観に基

表 2　カテゴリの設定の例
Table 2　Example of categories in Foursquare and Google.

Foursquare （）内は意味		Google
4fceea171983d5d06c3e9823	(Aquarium)	Aquarium
4d4b7104d754a06370d81259	(Art & entertainment)	Art gallery
4deefb944765f83613cdba6e	(Historic site)	Political
4bf58dd8d48988d181941735	(Museum)	Museum
4bf58dd8d48988d182941735	(Theme park)	
4bf58dd8d48988d1df941735	(Bridge)	
4bf58dd8d48988d163941735	(Park)	Park
4bf58dd8d48988d161941735	(Lake)	Place of worship
4bf58dd8d48988d129941735	(City hall)	City hall
4bf58dd8d48988d1f9931735	(Road)	Sublocality
4bf58dd8d48988d12d941735	(Monument landmark)	Establishment
4bf58dd8d48988d17b941735	(Zoo)	Zoo
4bf58dd8d48988d164941735	(Plaza)	Neighborhood
4d954b16a243a5684b65b473	(Rest area)	
4bf58dd8d48988d129951735	(Train station)	
5032792091d4c4b30a586d5c	(Concert hall)	
4bf58dd8d48988d15a941735	(Garden)	
4bf58dd8d48988d184941735	(Stadium)	
4bf58dd8d48988d132941735	(Charch)	Church
4bf58dd8d48988d1f2931735	(Performing arts venue)	
4bf58dd8d48988d1fa941735	(Farmers Market)	

表 3　カテゴリ設定の効果（Foursquare API の場合）
Table 3　Effect of category filtering (Foursquare API).

クラスタの中心座標		第 1 候補（カテゴリ設定なし）		第 1 候補（カテゴリ設定あり）	
緯度	経度	名前	カテゴリ	名前	カテゴリ
40.759001	-73.979122	30 Rockefeller Plaza	Building	Rockefeller Center	Plaza
40.757847	-73.985673	Microsoft Pop-Up Store	Electronics Store	Discovery Times Square	Museum
40.741607	-73.989340	The Flatiron District	Neighborhood	Flatiron Building	Historic Site
37.762226	-122.435068	Hot Cookie	Bakery	Castro Theater	Indie Movie Theater
37.808730	-122.415708	The Chowder Hut	Seafood Restaurant	Fisherman's Wharf Sign	Historic Site

づいてカテゴリを設定し，カテゴリ指定の有無によって結果に差が出るかを検証した．

提案システムでは，「観光」に関する情報を抽出することを目的としているため，**表 2** に示すように，Foursquare API と Google API に対して，それぞれ 21 個と 12 個のカテゴリを設定した．

その結果の一部を**表 3** に示す．カテゴリを指定しない場合には Bekery や Seafood Restaurant が第 1 候補として表示されていた位置に対して，カテゴリを指定した場合，Movie Theater や Historic Site など，観光に関係しそうな POI が第 1 候補として選出されており，一定の効果を確認できる．

将来的には，ユーザの挙動（提示された POI に対するクリックなど）に応じて，目的に対するカテゴリのセットを自動形成する仕組みを検討していきたいと考えている．

4.2　データセットのサイズに関する事前実験

Mean Shift 法を用いてクラスタリングを行う場合，データセットのサイズが小さいほど，計算時間が短くなるのは自明である．一方，データセットを小さくすると，抽出された結果の信頼性が低下する可能性がある．また，単一の撮影者が同じ場所で同じ時間帯に連射すると分析に影響を与えてしまうことも自明である．そこで本研究では，分析に十分なデータセットのサイズについて調査する．事前実験では，ロンドンの 1.9 km 四方エリア[*14]とパリの 3.77 km 四方エリア[*15]を対象として，収集したデータの中からランダムに，1 万枚，5 万枚，10 万枚，30 万枚を抽出して，4 通りのデータセットを作成し，それぞれに対して Bandwidth を 0.001（100 m）として Mean Shift 法によるクラスタリングを行い，含まれる写真の数が多い上位 10 クラスタとその中心点の座標を比較する．さらに正解値として，各クラスタの中心点およびタグ分析結果に基づいて人為的に決定された POI 名とその座標を示す．このとき POI の座標は，Wikipedia に登録されている座標を用いる．

図 1 に，ロンドンにおいて 4 通りのデータセットを用いて Mean Shift 法を適用した結果を示す．ここで，Bandwidth

*14 Google Static Maps で Zoom レベルを 15 として 600 px 四方で切り出した場合の実距離
*15 Google Static Maps で Zoom レベルを 14 として 600 px 四方で切り出した場合の実距離

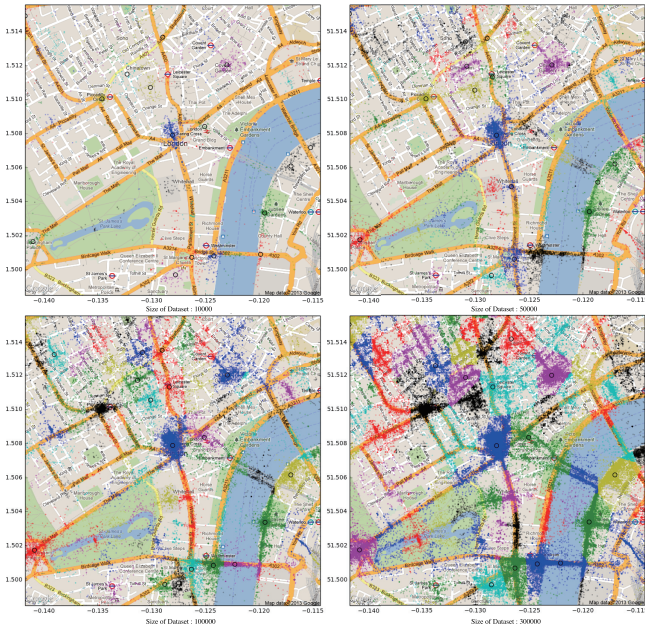

図 1　ロンドンの結果（それぞれ約 1.9 km 四方のエリア）
Fig. 1　Clustering result at London (1.9 km square meters area).

あたりのデータの密度を表す DPB（Data Per Bandwidth）という指標を導入する．

$$DPB = \frac{The\ size\ of\ dataset}{\left(\frac{One\ side\ length\ of\ the\ area\ (m)}{Actual\ distance\ for\ Bandwidth\ (m)}\right)^2} \qquad (4)$$

たとえば，ロンドンの場合，1 辺は 1.9 km であるため，式 (4) を用い，10,000 枚の写真データを利用する場合，その DPB は 27.7 と算出できる．図 1 を見ると，主観的には，10 万枚のデータセット（DPB: 277）と 30 万枚のデータセット（DPB: 831）の結果は，見た目上，あまり変化がないように見える．一方，10,000 枚のデータセット（DPB: 27.7）はデータが不足しているように見える．次に，より詳しい結果を**表 4**，**表 5**，**表 6**，**表 7** に示す．まず，上位 2 件に関しては，どのデータセットを用いても同じ結果になっており，かつ，実際の位置との誤差はいずれも非常に小さいことが分かる．Buckingham Palace と St Paul's Cathedral については，データセットによって有無が異なるが，出現する場合もその位置の誤差はいずれも大きい．これは POI の物理的なサイズが大きいため，写真撮影地点（ジオタグに記録される位置）と実際の POI の位置が離れているからであると考えられる．これらの結果から，ロン

表 4　上位 10 件とその位置精度（Dateset Size = 10,000）
Table 4　Top 10 spots with location accuracy (dataset size = 10,000).

	名前	Wikipedia		クラスタリングの中心		誤差 (m)
		緯度	経度	緯度	経度	
1	Trafalgar Square	51.508056	-0.128056	51.507906	-0.128039	16.79
2	The London Eye	51.5033	-0.1197	51.503323	-0.119459	16.91
3	British Museum	51.519459	-0.126931	51.519250	-0.126861	23.8
4	Tate Modern	51.507778	-0.099167	51.507880	-0.099232	12.24
5	Covent Garden	51.51197	-0.1228	51.512034	-0.122969	13.74
6	Piccadilly Circus	51.51	-0.134444	51.510038	-0.134596	11.36
7	Royal Festival Hall	51.505836	-0.116789	51.506088	-0.117014	32.05
8	Big Ben	51.500756	-0.124661	51.500796	-0.124192	32.89
9	Buckingham Palace	51.501	-0.142	51.501642	-0.141013	98.99
10	Parliament Square	51.500556	-0.126667	51.500711	-0.126233	34.69

表 5　上位 10 件とその位置精度（Dateset Size = 50,000）
Table 5　Top 10 spots with location accuracy (dataset size = 50,000).

	名前	Wikipedia		クラスタリングの中心		誤差 (m)
		緯度	経度	緯度	経度	
1	Trafalgar Square	51.508056	-0.128056	51.507880	-0.128022	19.70
2	The London Eye	51.5033	-0.1197	51.503378	-0.119408	22.02
3	British Museum	51.519459	-0.126931	51.519267	-0.126871	21.78
4	Tate Modern	51.507778	-0.099167	51.507894	-0.099231	13.6
5	Covent Garden	51.51197	-0.1228	51.512025	-0.122886	8.53
6	Piccadilly Circus	51.51	-0.134444	51.510031	-0.134571	9.47
7	Big Ben	51.500756	-0.124661	51.500851	-0.124221	32.38
8	Parliament Square	51.500556	-0.126667	51.500578	-0.126369	20.83
9	Royal Festival Hall	51.505836	-0.116789	51.506191	-0.117029	42.82
10	Buckingham Palace	51.501	-0.142	51.501746	-0.140817	116.79

表 6　上位 10 件とその位置精度（Dateset Size = 100,000）
Table 6　Top 10 spots with location accuracy (dataset size = 100,000).

	名前	Wikipedia		クラスタリングの中心		誤差 (m)
		緯度	経度	緯度	経度	
1	Trafalgar Square	51.508056	-0.128056	51.507883	-0.128019	19.46
2	The London Eye	51.5033	-0.1197	51.503369	-0.119429	20.35
3	Tate Modern	51.507778	-0.099167	51.507891	-0.099228	13.26
4	St Paul's Cathedral	51.513611	-0.098056	51.513790	-0.099014	69.4
5	British Museum	51.519459	-0.126931	51.519275	-0.126882	20.74
6	Royal Festival Hall	51.505836	-0.116789	51.506154	-0.117030	39.14
7	Piccadilly Circus	51.51	-0.134444	51.510037	-0.134581	10.37
8	Covent Garden	51.51197	-0.1228	51.512018	-0.122929	10.42
9	Big Ben	51.500756	-0.124661	51.500840	-0.124235	31.05
10	Buckingham Palace	51.501	-0.142	51.501731	-0.140867	113.12

表 7　上位 10 件とその位置精度（Dateset Size = 300,000）
Table 7　Top 10 spots with location accuracy (dataset size = 300,000).

	名前	Wikipedia		クラスタリングの中心		誤差 (m)
		緯度	経度	緯度	経度	
1	Trafalgar Square	51.508056	-0.128056	51.507875	-0.128016	20.3
2	The London Eye	51.5033	-0.1197	51.503375	-0.119425	20.87
3	Tate Modern	51.507778	-0.099167	51.507879	-0.099258	12.94
4	British Museum	51.519459	-0.126931	51.519267	-0.126877	21.67
5	Covent Garden	51.51197	-0.1228	51.512022	-0.122907	9.39
6	Royal Festival Hall	51.505836	-0.116789	51.506151	-0.117030	38.88
7	Piccadilly Circus	51.51	-0.134444	51.510033	-0.134571	9.57
8	Big Ben	51.500756	-0.124661	51.500848	-0.124237	31.14
9	Parliament Square	51.500556	-0.126667	51.500605	-0.126310	25.38
10	St Paul's Cathedral	51.513611	-0.098056	51.513673	-0.098283	17.18

ドンに関しては，ランダムサンプリングによって得られた 10,000 件のデータセットでも，30 倍のデータセットと遜色ない結果が得られることが分かる．

次に，ロンドンよりもデータセットあたりの面積を大きく設定したパリについて分析した結果を**図 2** に示す．図中の各地図の 1 辺は約 3.8 km に相当する．そのため，各データセットの DPB は，小さい順に，それぞれ 6.9, 34.6, 69.3, 207.8 となる．DPB がきわめて小さい 10,000 枚のデータセットの場合，クラスタと呼べるものが少なく，DPB が増加するに従い，クラスタが鮮明になることが分かる．ロンドンと同様に各データセットにおける上位 10 件の詳しい結果を**表 8**，**表 9**，**表 10**，**表 11** に示す．

結論から述べると，予想外に，低い DPB の場合も，高い DPB の場合とほぼ同じ 10 件の POI を抽出でき，その位置誤差も小さいことが分かる．Effel Tower と Louvre Pyramid に注目すると，その順位はデータセットによって異なるが，その位置誤差はどのデータセットでも同等（Effel Tower は約 11 m，Louvre Pyramid は約 22 m）であることが分かる．この評価における順位は，クラスタ内の写真の枚数に基づいているため，ランダムに抽出した過程で，誤差に影響を与えない程度のわずかな枚数の差だけが生じたと予想される．筆者らは，枚数だけでなく，時間分

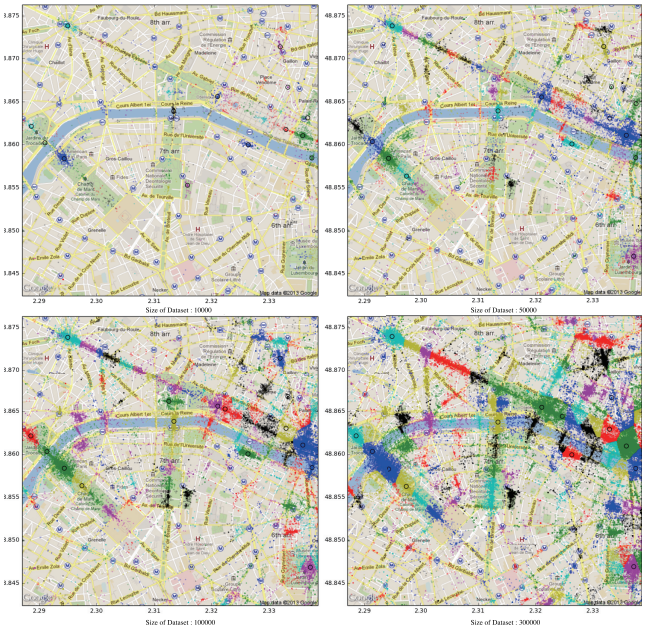

図 2　パリの結果（それぞれ約 3.7 km 四方のエリア）
Fig. 2　Clustering result at Paris (3.7 km square meters area).

表 8　上位 10 件とその位置精度（Dateset Size = 10,000）
Table 8　Top 10 spots with location accuracy (dataset size = 10,000).

	名前	Wikipedia		クラスタリングの中心		誤差 (m)
		緯度	経度	緯度	経度	
1	Eiffel Tower	48.8583	2.2945	48.858367	2.294374	11.89
2	Louvre Pyramid	48.860854	2.335812	48.861042	2.335898	21.83
3	Notre Dame de Paris	48.853	2.3498	48.853165	2.349368	36.65
4	Arc de Triomphe	48.8738	2.295	48.873828	2.294993	3.11
5	Pompidou Centre	48.860653	2.352411	48.860528	2.352117	25.69
6	Basilique du Sacré-Cœur	48.886694	2.343	48.886264	2.343023	47.85
7	Place de l'Hôtel de Ville	48.856667	2.351389	48.856748	2.351397	8.99
8	Musée d'Orsay	48.86	2.327	48.859987	2.326399	44.14
9	Pont des Arts	48.858333	2.3375	48.858448	2.337483	12.87
10	Musée du Louvre	48.860339	2.337599	48.860459	2.339875	167.51

表 9　上位 10 件とその位置精度（Dateset Size = 50,000）
Table 9　Top 10 spots with location accuracy (dataset size = 50,000).

	名前	Wikipedia 緯度	Wikipedia 経度	クラスタリングの中心 緯度	クラスタリングの中心 経度	誤差 (m)
1	Louvre Pyramid	48.860854	2.335812	48.861050	2.335913	22.99
2	Eiffel Tower	48.8583	2.2945	48.858364	2.294393	10.58
3	Notre Dame de Paris	48.853	2.3498	48.853160	2.349361	36.80
4	Arc de Triomphe	48.8738	2.295	48.873831	2.294998	3.44
5	Pompidou Centre	48.860653	2.352411	48.860540	2.352162	22.17
6	Basilique du Sacré-Cœur	48.886694	2.343	48.886287	2.343054	45.42
7	Place de l'Hôtel de Ville	48.856667	2.351389	48.856702	2.351540	11.74
8	Pont des Arts	48.858333	2.3375	48.858437	2.337526	11.68
9	Musée d'Orsay	48.86	2.327	48.860027	2.326363	46.87
10	Pont Neuf	48.857447	2.341617	48.857128	2.341052	54.58

表 10　上位 10 件とその位置精度（Dateset Size = 100,000）
Table 10　Top 10 spots with location accuracy (dataset size = 100,000).

	名前	Wikipedia 緯度	Wikipedia 経度	クラスタリングの中心 緯度	クラスタリングの中心 経度	誤差 (m)
1	Louvre Pyramid	48.860854	2.335812	48.861046	2.335894	22.15
2	Eiffel Tower	48.8583	2.2945	48.858364	2.294392	10.62
3	Notre Dame de Paris	48.853	2.3498	48.853158	2.349346	37.65
4	Arc de Triomphe	48.8738	2.295	48.873819	2.295016	2.36
5	Pompidou Centre	48.860653	2.352411	48.860523	2.352174	22.62
6	Basilique du Sacré-Cœur	48.886694	2.343	48.886297	2.343051	44.30
7	Place de l'Hôtel de Ville	48.856667	2.351389	48.856706	2.351574	14.29
8	Place de la Concorde	48.865556	2.321111	48.865512	2.321118	4.93
9	Pont des Arts	48.858333	2.3375	48.858427	2.337527	10.64
10	Pont Neuf	48.857447	2.341617	48.857191	2.340981	54.71

表 11　上位 10 件とその位置精度（Dateset Size = 300,000）
Table 11　Top 10 spots with location accuracy (dataset size = 300,000).

	名前	Wikipedia 緯度	Wikipedia 経度	クラスタリングの中心 緯度	クラスタリングの中心 経度	誤差 (m)
1	Eiffel Tower	48.8583	2.2945	48.858363	2.294394	10.52
2	Louvre Pyramid	48.860854	2.335812	48.861049	2.335900	22.61
3	Notre Dame de Paris	48.853	2.3498	48.853157	2.349369	36.16
4	Arc de Triomphe	48.8738	2.295	48.873828	2.294996	3.18
5	Pompidou Centre	48.860653	2.352411	48.860536	2.352158	22.70
6	Basilique du Sacré-Cœur	48.886694	2.343	48.886303	2.343056	43.65
7	Place de l'Hôtel de Ville	48.856667	2.351389	48.856718	2.351355	6.21
8	Pont des Arts	48.858333	2.3375	48.858436	2.337533	11.74
9	Place de la Concorde	48.865556	2.321111	48.865570	2.321133	2.25
10	Musée d'Orsay	48.86	2.327	48.860026	2.326380	45.61

散を加味した順位付けを行うことで，これらの順位誤差も低減させることができるのではないかと考えている．

5.　提案手法

本論文で提案する，ソーシャル観光マップは，位置情報付きのソーシャルデータの分析による都市の人気スポットを抽出して地図上に可視化するシステムであり，**図 3** に示すような構成となる．分析対象となるデータの情報源として，Flickr 上の位置情報付き写真を利用し，Mean Shift 法を用いてクラスタリングし，人気スポットを抽出するという全体の流れは，従来研究 [5], [6] と共通である．異なる点は，網掛けされた部分であり，計算の高速化を目的としたデータセットのサンプリング，チェックインサービスからの情報を統合した POI 名の推定手法，そして，枚数と撮影時間の時間分散を考慮した人気度の定量化である．なお，本論文では，副題のとおり，人気スポットの抽出に焦点を当てており，地図上に可視化するシステムに関しては今後の研究課題とする．

5.1　データセットのサンプリングに関して

今回，5 都市（New York, San Francisco, London, Paris, Berlin）で撮影された位置情報付き写真 436 万枚を Flickr から収集した．436 万枚の写真の撮影者は 15.4 万人にのぼり，撮影者あたりの写真の枚数は，28.4 枚となる．近年はデジタルカメラのメモリも大容量かつ安価になっているため，1 撮影者が連射で何枚も撮影していることも多い．そこで，従来研究と同様に，30 分以内に同じ撮影者によって撮影されたすべての写真を 1 つと見なす前処理を行う．提案では，古い写真を排除する（2004/01/01 00:00:00 以降の写真に限定する）．同時に，付与されている位置情報の精度が低い写真と，タグがいっさい付与されていない写真も候

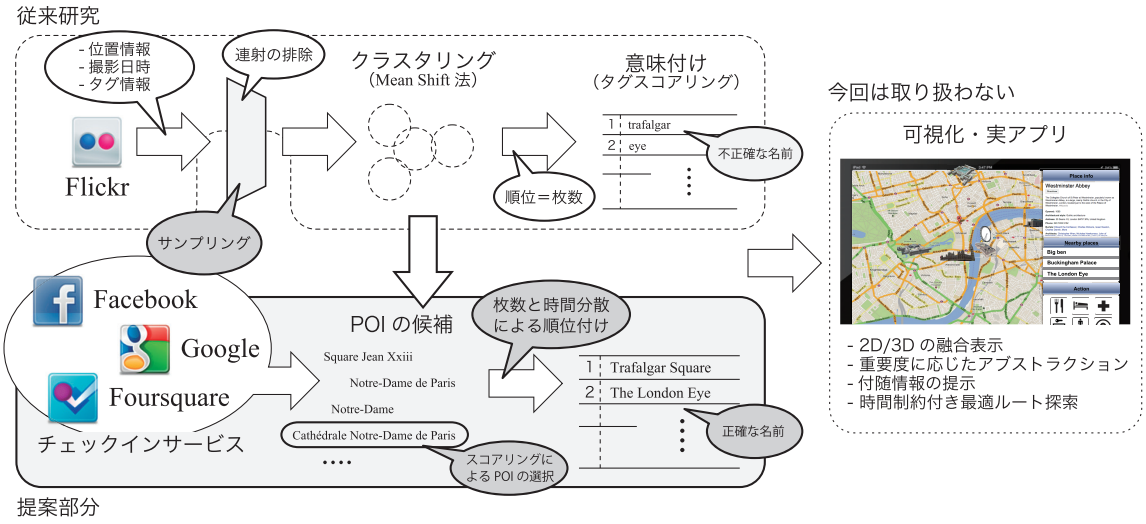

図 3　提案システムの構成と本論文で取り扱う項目
Fig. 3　Whole system architecture and a target of this paper.

表 12　ニューヨークにおける第 5 位のクラスタにおけるスコア計算の例
Table 12　Calculation example of the score for the 5th cluster of New York.

サービス名	出力順位	POI 名	平均編集距離によるスコア	出現頻度によるスコア	スコア
Foursquare	1 位	Museum of Modern Art (MoMA)	0.578109941	1.75	1.011692397
	2 位	The Paley Center for Media	0.484803086	0	0
	3 位	Saint Thomas Church	0.451673654	0	0
Facebook	1 位	The Metropolitan Museum of Art	0.593139804	1.333333333	0.790853072
	2 位	MoMA	0.413197612	1	0.206598806
	3 位	The Modern - Dining Room	0.563627745	1	0.187875915
Google	1 位	Museum of Modern Art	0.587213362	2.333333333	1.370164512
	2 位	21 Club	0.337940269	0	0
	3 位	The Modern	0.62388356	3	0.62388356

補から除く．その結果，分析対象となるデータは，182 万枚に絞り込まれる．さらに，提案システムでは，事前実験の検証結果に基づき，この絞り込まれたデータから，さらにランダムサンプリングすることで所望のサイズのデータセットを作成する．今回は上位 10 件だけに焦点を当てることから，DPB が 20 以上となるデータセット（New York: 200,000，San Francisco: 300,000，London: 20,000，Paris: 50,000，Berlin: 100,000）を用いる．サンフランシスコは対象となるエリアが大きいため，より多くのデータが必要となる．一方，ロンドンは最もエリアが狭く，小さなデータセットで DPB が 20 以上となる．

5.2　チェックインサービスの統合に関して

今回，3 つのチェックインサービス（Foursquare，Facebook，Google）が提供しているリバースジオコーディング API を用いる．Foursquare と Google に関しては，事前実験の検証結果に基づき，観光に関するカテゴリ設定を行う．また，Google は距離に基づいた出力も可能であるが，今回は他と合わせるために，重要度に基づいた出力を指定する．あるクラスタの中心座標（x）として，リバースジオコーディング API から得られる上位 m の POI 名 $\{s_1, s_2, \cdots, s_m\}$ のうちから最も確からしい s を選択する手法について考える．

提案手法では，確からしさを「他の候補との類似性」と「単語の出現頻度」という 2 つの指標で評価する．他の候補との類似性は，文字列間の編集距離を計算し，他の $m-1$ 個の POI との平均編集距離 d_m を求める．編集距離の計算は，有名な Levenshtein 距離でもよいが，今回は扱いやすさの観点[*16]から Jaro-Winkler 距離 [15] を利用している．単語の出現頻度は，$s_{i\ (i=\{1,2,\cdots,m\})}$ をさらに n 個の単語 $w_1^i, w_2^i, \cdots, w_n^i$ に分割し，各単語がそれぞれ何回ほかの POI 名で利用されているかを各単語の重みとし，その総和を含まれる単語数で除算したものを POI 名 s の出現頻度によるスコアとする．単語数で除算する理由は，POI 名の長さの影響を緩和するためである．また，the や of や記号などのストップワードは，単語と見なさず，すべて重みを 0 とする．これに先ほど計算した d_m を乗算し，出現順位で除算したものを POI 名 $s_{i\ (i=\{1,2,\cdots,m\})}$ の最終的なスコアとし，そのスコアが大きなものを最も確からしい POI 名として選出する．出現順位で除算するのは，各 API で考慮されている人気度を反映するためである．提案アルゴリズムにより，チェックインサービスにおける人気度が高い POI の中で，多くの候補に含まれる単語を含みつつ，文字列全体に見たときに類似度の高い他の候補が存在するような POI が選ばれる．なお今回，3 つの API からそれぞれ上位 3 件を候補としているため，m は 9 となる．

表 12 に，ニューヨークにおける第 5 位のクラスタに関する，スコアの数値例を示す．各 API からの出力結果から，正解となる POI 名は「Museum of Modern Art（ニューヨーク近代美術館）」と推測できるが，その表記は API によってさまざまであることが分かる．この中で，最も他の候補との類似度が高い（平均編集距離によるスコアが高い）のは，Google API の 3 位として得られた「The Modern」である．また，この中の「Modern」という単語は，他にも 3 つの候補で利用されており，その重みは 3 となる．そして，The Modern に含まれる単語数は，ストップワードである The を除外するため 1 となり，出現頻度によるスコアは 3 と計算できる．しかしながら，Google における順位が 3 位であるため，最終的なスコアはそれほど大きな値にはならない．最終スコアが最も高くなったのは，Google API の 1 位として得られた「Museum of Modern Art」である．平均編集距離によるスコアは全体の 3 位，出現頻度によるスコアは全体の 2 位だが，Google における順位は 1 位であり，最終的なスコアは大きな値となる．このように提案アルゴリズムは，各 API における出力順位が大きく影響する．これは，アルゴリズムは不明であるものの，各社における膨大なデータを用いた人気度計算を重視しているためである．ちなみに，この例において，従来のタグ分析によって得られた POI 名は，museumofmodernart，であり，提案手法により，適切かつ正確性の高い POI 名が選出できていることが分かる．

[*16] Jaro-Winkler 距離は 0〜1 の値となるが，Levenshtein 距離は文字列長によって最大値が異なる．それを正規化する手法も提案されているが，今回は Jaro-Winkler 距離を用いる．

表 13　上位 10 件とその名前に関する比較（ロンドン）
Table 13　Comparison of top 10 spots and their names (London).

順位	従来方式	提案方式
1	trafalgar	Trafalgar Square
2	tatemodern	The London Eye
3	britishmuseum	St Paul's Cathedral
4	eye	Tate Modern
5	stpaulscathedral	British Museum
6	covent	Big Ben
7	royalfestivalhall	Piccadilly Circus
8	parliamentsquare	Parliament Square
9	bigben	Covent garden
10	piccadillycircus	Buckingham Palace Gardens

5.3　時間分散を考慮した人気度について

本研究は，観光スポットの抽出を目的としているため，定常的に人気度の高いスポットを抽出する仕組みが必要である．従来方式では，単にクラスタ内の写真の枚数によってクラスタを順位付けしていたが，この手法はジオタグ付き写真がたまたま多く発生した大きなイベントの影響を受けることがある．また，わずか数枚の写真枚数の違いでスポットの人気度の順位が変わるのも意にそぐわない．

本論文では，有名な観光スポットは今も昔も有名という前提に基づき，写真が定常的に撮影されているか否かによって，そのスポットの観光という目的に対する重要度を決定する仕組みを提案する．定常性を測るために，本論文では，クラスタ内の写真をタイムスタンプ順にソートし，写真の撮影間隔の分散を計算する．クラスタ c に k 枚の写真が含まれてているとしたとき，古い順にソートしたタイムスタンプ群を p_i（$i = 1, \cdots, k$）と定義する．最古のタイムスタンプは p_1，最新のタイムスタンプは p_k となる．このとき，写真の撮影間隔 W_i は $W_i = p_i - p_{(i-1)}$（$i = \{0, \cdots, k\}$）と表すことができる．p_0 は，データセットに含まれる可能性のある最も古いタイムスタンプ 2004/01/01 00:00:00 とする．この W_i を用いて，クラスタ c に含まれる写真の撮影時間の分散 D_c は，$D_c = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(W_i - \bar{W})^2}$ と計算することができる．提案手法では，この D_c にクラスタ内の写真の枚数を乗算した，$D_c \times k$ をクラスタ c の重要度と定義する．

6.　分析結果

今回，データを収集した 5 都市に関して，従来方式（枚数による順位付け＋タグ分析による意味付け）と提案方式（枚数と時間分散による順付け＋チェックインサービスを用いた意味付け）による観光スポット上位 10 件の比較を行う．このとき，データセットのサイズは，事前実験の結果に基づき，それぞれ異なるサイズを用いている．

表 13，**表 14**，**表 15**，**表 16**，**表 17** の結果を見ると，いずれも提案手法によって，正確性の高い名前が割り当てできていることが分かる．しかしながら，その順位は，あまり大きな違いは見られない．また，順位の入れ替わりが，本当に人気度を示しているのかは今回の評価では評価できていないため不明であり，今後，アプリケーションをリリースし，ユーザスタディを通じて，順位付けの評価を行いたいと考えている．

表 14　上位 10 件とその名前に関する比較（サンフランシスコ）
Table 14　Comparison of top 10 spots and their names (SF).

順位	従来方式	提案方式
1	unionsquare	Alcatraz Island
2	prison	Coit Tower
3	attpark	San Francisco City Hall
4	californiaacademyofsciences	Union Square
5	cityhall	Sea Lions @ Pier 39
6	sfmoma	Powell St. BART Station
7	flickrhq	Ferry Building Marketplace
8	sanfrancisco	de Young Museum
9	ferrybuilding	San Francisco Museum of Modern Art
10	deyoungmuseum	Transamerica Redwood Park

表 15　上位 10 件とその名前に関する比較（ニューヨーク）
Table 15　Comparison of top 10 spots and their names (New York).

順位	従来方式	提案方式
1	rockefellercenter	Rockefeller Center
2	timessquare	Empire State Building
3	empirestatebuilding	Prayer in the Square
4	museumofmodernart	Times Square
5	timessquare	Museum of Modern Art
6	grandcentralterminal	Flatiron Building
7	flatironbuilding	Grand Central Terminal
8	bryantpark	Wall Street
9	—	Bryant Park
10	unionsquare	Washington Square Park

表 16　上位 10 件とその名前に関する比較（パリ）
Table 16　Comparison of top 10 spots and their names (Paris).

順位	従来方式	提案方式
1	pyramid	Cathédrale Notre-Dame de Paris
2	notredame	Tour Eiffel
3	eiffeltower	Pyramide du Louvre
4	centrepompidou	Centre Pompidou - Musée National d'Art Moderne
5	sacrecoeur	Arc de Triomphe
6	arcdetriomphe	Musée d'Orsay
7	Paris	Square Jean XXIII
8	pontdesarts	Pont des Arts
9	saintechapelle	Sainte Chapelle
10	placedelaconcorde	Place de la Concorde

表 17　上位 10 件とその名前に関する比較（ベルリン）
Table 17　Comparison of top 10 spots and their names (Berlin).

順位	従来方式	提案方式
1	pariserplatz	Brandenburg Gate
2	reichstag	Reichstag
3	potsdamer	Potsdamer Platz
4	holocaustmemorial	CineStar Sony Center
5	alexanderplatz	Alexanderplatz
6	sonycenter	Holocaust Mahnmal
7	—	Checkpoint Charlie
8	berlinhauptbahnhof	S+U Bahnhof Berlin Alexanderplatz
9	berlinerdom	Berliner Dom
10	deutscherdom	Berlin Brandenburger Tor station

7.　おわりに

本論文では，位置情報付きのソーシャルデータを分析に

基づくソーシャル観光マップの構築に向け，都市の人気スポットをその正確な名前とともに抽出する仕組みを提案した．Foursquare などの複数のチェックインサービスから得られる情報を用いる提案手法によって，従来のタグ分析手法と比較して，より正確な表記の名前を得られることを明らかにした．また，提案手法を用いることにより，小さなデータセットであっても，正確性の高い意味付けが可能となり，データセットサイズの削減による計算速度の改善が見込めることを明らかにした．

謝辞　本研究の一部は，総務省戦略的情報通信研究開発推進制度（SCOPE）の支援を受けて実施している．

参考文献

[1] 荒川　豊，末松慎司，田頭茂明，福田　晃：コンテキストアウェア IME の実現へ向けた動的辞書生成手法の提案，情報処理学会論文誌，Vol.52, No.3, pp.1033–1044 (2011).
[2] Wakamiya, S., Lee, R. and Sumiya, K.: Crowd-based urban characterization: extracting crowd behavioral patterns in urban areas from twitter, *Proc. 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, pp.77–84, ACM (2011).
[3] Ishikawa, S., Arakawa, Y., Tagashira, S. and Fukuda, A.: Hot topic detection in local areas using Twitter and Wikipedia, *ARCS Workshops* (*ARCS*), *2012*, pp.1–5, IEEE (2012).
[4] Chen, W., Battestini, A., Gelfand, N. and Setlur, V.: Visual summaries of popular landmarks from community photo collections, *2009 Conference Record of the 43rd Asilomar Conference on Signals, Systems and Computers*, pp.1248–1255, IEEE (2009).
[5] Crandall, D., Backstrom, L., Huttenlocher, D. and Kleinberg, J.: Mapping the world's photos, *Proc. 18th international conference on World wide web*, pp.761–770, ACM (2009).
[6] Kurashima, T., Iwata, T., Irie, G. and Fujimura, K.: Travel route recommendation using geotags in photo sharing sites, *Proc. 19th ACM international conference on Information and knowledge management*, pp.579–588 (2010).
[7] Yin, Z., Cao, L., Han, J., Luo, J. and Huang, T.: Diversified trajectory pattern ranking in geo-tagged social media, *Proc. 11th SIAM International Conference on Data Mining, SDM 2011*, pp.980–991 (2011).
[8] Liu, H., Wei, L.-Y., Zheng, Y., Schneider, M. and Peng, W.-C.: Route discovery from mining uncertain trajectories, *2011 IEEE 11th International Conference on Data Mining Workshops* (*ICDMW*), pp.1239–1242, IEEE (2011).
[9] Lowe, D.G.: Distinctive image features from scale-invariant keypoints, *International journal of computer vision*, Vol.60, No.2, pp.91–110 (2004).
[10] Cheng, Y.: Mean shift, mode seeking and clustering, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.17, No.8, pp.790–799 (1995).
[11] Carreira-Perpinan, M.: Acceleration strategies for Gaussian mean-shift image segmentation, *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol.1, pp.1160–1167, IEEE (2006).
[12] Kisilevich, S., Mansmann, F. and Keim, D.: P-DBSCAN: A density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos, *Proc. 1st International Conference and Exhibition on Computing for Geospatial Research & Application*, p.38, ACM (2010).
[13] Yang, Y., Gong, Z., et al.: Identifying points of interest by self-tuning clustering, *Proc. 34th international ACM SIGIR conference on Research and development in Information*, pp.883–892, ACM (2011).
[14] De Choudhury, M., Feldman, M., Amer-Yahia, S., Golbandi, N., Lempel, R. and Yu, C.: Automatic construction of travel itineraries using social breadcrumbs, *Proc. 21st ACM conference on Hypertext and hypermedia*, pp.35–44 (2010).
[15] Jaro, M.: Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida, *Journal of the American Statistical Association*, Vol.84, No.406, pp.414–420 (1989).

荒川 豊　（正会員）

1977 年生．2001 年慶應義塾大学理工学部情報工学科卒業．2003 年同大学大学院修士課程修了．2006 年同大学院博士課程修了．博士（工学）．2006 年同大学院特別研究助手（2007 年より助教に変更）．2009 年 3 月九州大学大学院システム情報科学研究院助教．2011 年 11 月 ENSEEIHT（Toulouse，France）訪問研究員．2012 年 2 月 DFKI（Kaiserslautern，Germany）訪問研究員．2013 年 3 月より奈良先端科学技術大学院大学准教授．主として，ネットワークアプリケーション，ソーシャルデータマイニング，に関する研究に従事．APCC 2008 Best Paper Award（2008），MBL 研究会優秀論文賞（2009，2011，2013），DICOMO 優秀論文賞（2010，2013），DICOMO 優秀プレゼンテーション賞（2010），山下記念研究賞（2011），安藤博記念学術奨励賞（2011），DPSWS 優秀論文賞（2012），DPSWS 優秀ポスター賞（2011，2013），DPSWS ベストカンバサント賞（2013），ICMU2014 Best Poster Award（2014），等受賞．IEEE，ACM，電子情報通信学会各会員．

Tatjana Scheffler

She studied computational linguistics at the University of the Saarland, Germany, and received a Ph.D. in linguistics from the University of Pennsylvania, U.S.A., in 2008. From 2008–2012, she was a researcher at the German Research Center for Artificial Intelligence (DFKI) in Berlin. She now works at the University of Potsdam, Germany. Her current research interests are discourse structure, natural language semantics, and social media processing.

Stephan Baumann

He heads the Competence Center Computational Culture (C4) at the German Research Center for AI in Kaiserslautern and Berlin (DFKI). He reveived the Ph.D. degree on Artificial Listening Systems at DFKI and IRCAM/Paris. His current research interests are in algorithm design for Social Network Analysis, Semantic Recommenders and the Post-Digital/Neo-Analog world. His research team at C4 works on realtime processing and datamining of large-scale social and sensor data.

Andreas Dengel

He is a member of the Management Board as well as Scientific Director at the German Research Center for Artificial Intelligence (DFKI) in Kaiserslautern where he is leading the Knowledge Management Research Department. In 1993 he became a Professor at the Computer Science Department of the University of Kaiserslautern. Since 2009 he is also appointed Professor (Kyakuin) at the Dept. of Computer Science and Intelligent Systems, Graduate School of Engineering of the Osaka Prefecture University. From 1980 to 1986, he studied Computer Science and Economics at the University of Kaiserslautern. He subsequently worked at the Siemens research lab in Munich and at the University of Stuttgart where he completed his doctoral thesis in 1989. In 1991 he worked as a guest researcher at Xerox Parc in Palo Alto. He is co-editor of various international computer science journals and has written or edited 11 books and is author of more than 240 peer-reviewed scientific publications, some of which received a Best-Paper Award.