ソーシャルネットワークを用いた実世界センシングの研究

研究代表者　　荒川　豊　　奈良先端科学技術大学院大学　情報科学研究科　准教授

1 はじめに

近年，Facebook や Twitter などに代表されるソーシャルネットワークサービス（SNS: Social Network Service）の普及が著しい．今や Twitter はユーザ数 5 億人を抱え，3 日で 10 億ツイートもの情報が発信されている．Facebook はユーザ数 10 億人に迫っており，1 日当たり 32 億コメント，3 億枚の写真投稿が行われている．そして，この膨大なデータから，新たなインテリジェンスを得ようとする研究が盛んに行われており，特に，研究代表者は，位置情報を含んだデータに着目して研究を進めている．例えば，2010 年に発表した「位置連携日本語入力システム GeoIME[1]」では，位置に応じて文字変換に用いる辞書が更新されていくものであり，この有効性を検証するために，Twitter 上の文字列分析を行なっている[2].

今回，貴財団の援助により 1 年間滞在したドイツ人工知能研究所では，「観光（都市分析）」に焦点をあて，ソーシャルネットワーク上のデータを分析することによって，実世界でどんなイベントが起きているのか，また実世界のどんなオブジェクトが人を惹きつけているのか，などを明らかにする手法について研究を行った．特に，位置データのマイニング結果に対して，別のソーシャルネットワークの情報を用いて意味付けする手法に関して研究を進め，従来のタグ（ソーシャルデータに付随する別の文字列情報）を用いる手法よりも正確な意味を割り当て可能であることを明らかにした．

2 研究概要

本研究は，ソーシャルネットワーク上のデータを分析して，さまざまな都市の観光地図を自動的に生成することが目的である．そのために，以下に示す３つの研究トピックに取り組んだ．

2-1 十分なデータセットサイズの調査

これまで我々は，Flickr からのべ 430 万枚以上（5 都市）の位置情報付き写真を収集している．最も多いエリアは，ニューヨークとサンフランシスコでそれぞれ約 110 万枚，次にロンドンが約 100 万枚，パリが 83 万枚，ベルリン 30 万枚である．これらのデータを分析するためには計算コストが高いため，分析する前に，適切なデータセットに絞り込むことが重要であると考えている．Crandall ら[1]は，単一の撮影者による同じ場所での連射の影響を排除するため，同一撮影者によって 30 分以内に撮影された写真を排除している．これに加えて，我々は，付与された位置情報の粒度（Flickr 上では，Accuracy が 1〜16 で定義されており，16 が最も細かい粒度を示す）と，写真に付随するタグの有無を考慮して，さらに絞り込む．さらに，その中からランダムに数万件〜数十万件をサンプリングし，それぞれの上位 10 クラスタの中心点の精度とそれが後述の意味付けにどのように影響するかを分析する．

2-2 チェックインとの紐付けによる POI 推定

従来の写真そのものに付与されたタグ情報を用いるのではなく，写真と同じく位置情報が付与されたデータであるチェックイン情報から，そのクラスタが何を意味しているのかを推定する手法を提案する．提案手法では，Foursquare，Facebook，Google という 3 大メジャーチェックインサービスを利用し，クラスタの中心点に対して提示される複数のチェックイン候補（アルゴリズムはブラックボックスであるが，各社ともに人気度に基づいた候補が提示されている）の中から，そのクラスタが示す POI を推定する．推定精度を改善するために，カテゴリによる候補の絞り込みをおこなった上で，各サービスから取得した上位３件のチェックイン候補，合計９件を，各候補間の文字列類似度，および，全候補中の単語の出現頻度による重み付けにより，順位付けすることによって，最も確からしい POI を推定する．

2-3 時間分散を考慮した人気度の定量化

「観光」は古くからあるものであり，人気観光スポットは今も昔も観光スポットである可能性が高い．そこ

で，各クラスタ内の写真の撮影時間に着目し，その撮影間隔が定常的であるほど，そのクラスタは定番観光スポットであると判断することができると考えている．提案手法では，写真の枚数に加えて，撮影間隔の分散を加味することで，観光スポットの定常的人気度を定量化する．

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，従来論文におけるタグ情報の順位付け手法，並びに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[2]は，主に画像分析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[1]からは，いくつかの研究[3][4]で空間情報のクラスタリングに用いられている．空間情報のクラスタリング手法としては，この Mean Shift 法以外にも，Kisilevich らによる p-DBSCAN[5]や，Yang らによる Self-tuning Spectral Clustering [6]などが提案されているが，これらが目的としている POI の大小や形状の違いを考慮する必要がないことから，パラメータが少ないという利点を重視し，本研究では 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．

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

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，従来研究において，一様カーネル[1]，またはガウシアンカーネル[3]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，と　$x_{(i+1)} = x_i + m_{h,G}(x_i)$　いう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．空間情報分析においては，Bandwidth w=0.001 は約 100m，w=1 は約 100km を表す．Crandall ら[1]は，全世界から都市を抽出する際に 1，各都市のスポットを抽出する際に 0.001 としている．他にも Yang ら[6]がスポットの抽出を目的として，0.001 としている．また，倉島ら[3]はルート分析と推薦に関する研究であるため，0.0001(10m)と極めて細かい粒度としているが，本研究では，スポットの抽出に相当するため，0.001 を用いる．

3-2 タグ情報の順位付け手法

Flickr の画像に付随するタグ情報の評価[1][3]について説明する．タグ情報の評価とは，Mean Shift 法によって生成された各クラスタに含まれるすべてのタグの中から，そのクラスタの特徴を表すタグを選出することである．選出にあたり，各タグ V のスコア $T(V)$ を下記の評価式によって求める．

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

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．この式により，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るため，タグスコアが大きいほどクラスタ m にのみよく現れるタグとなる．ただし，ノイズ（クラスタ特有であるが，スラングなど有用性の低い単語）を排除するため，各クラスタにおいてタグ V を含む写真の枚数が 5%以下のタグに関しては評価の対象から除く．

3-3 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伴い，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(Application Programming Interface) 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である．特に，「GeoNames (http://www.geonames.org/)」と「OpenStreetMap (http://www.openstreetmap.org/)」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，米 Foursquare 社（http://foursquare.com/）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，よりユーザの所望するチェックイン候補を上位に提示した方が利便性が向上することから，各社データベースおよび選出アルゴリズムを独自に構築している．本研究では，この各社が保有する POI データベースおよび選出アルゴリズムを利用する．これらのデータベースは，基本的には，取得したい位置情報（緯度・経度）と半径を指定すると独自の選出アルゴリズムによっていくつかのチェックイン候補が得られる．元となるデータセットは各社異なり，Foursquare であれば，上述した GeoNames のデータセット，Facebook であれば Factual 社の商用データセットが用いられている．選出アルゴリズムも各社で異なり，これは非公開である．他の相違点としては，ユーザが POI を追加可能か否かという点と，検索する際にカテゴリを指定可能か否かという点である．また，Foursquare だけのみ POI データをユーザ自身が追加することができる．これにより，他社と比較して膨大な登録データ数を誇るが，一方で表記揺らぎや表記ミスが多いという欠点も生じている．

4 事前実験

ここでは提案方式において，必要なパラメータを決定するために行った事前実験について説明する．今回は，「旅行」に関する情報のみを抽出することを考えていることから，まずカテゴリ設定が可能な API に対してカテゴリを設定することにより，提示される POI 候補の精度が改善すると考え，その効果について検証した．また，Mean Shift 法は，分析するデータセットのサイズにより計算時間が変化し，小さなデータセットであるほど高速に計算可能である．そこで，データセットのサイズが分析結果に与える影響について事前実験により調査した．

4-1 カテゴリ設定に関する事前実験

Foursquare API と Google API に対して，それぞれ表 1 に示す 21 個と 12 個のカテゴリを設定し，その設定の有無による結果の差を比較した結果の一部を表 2 に示す．ちなみに，一言でカテゴリと言っても，Foursquare と Google では抽象度も管理 ID も大きく異なっている．具体的には，Foursquare の場合，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．

カテゴリを指定しない場合には Bakery 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．

表 1　カテゴリの設定例

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 2　カテゴリ設定の効果

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
40.75900103	-73.9791215	30 Rockefeller Plaza	Building	Rockefeller Center	Plaza
40.75784677	-73.9856731	Microsoft Pop-Up Store	Electronics Store	Discovery Times Square	Museum
40.74160657	-73.98933982	The Flatiron District	Neighborhood	Flatiron Building	Historic Site
37.76222647	-122.4350683	Hot Cookie	Bakery	Castro Theater	Indie Movie Theater
37.80872975	-122.4157081	The Chowder Hut	Seafood Restaurant	Fisherman's Wharf Sign	Historic Site

今回は，「旅行」という目的に対して，著者が主観的にカテゴリを選択したが，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とする予定である．これにより，旅行以外の目的に対しても対応できると考えている．

4-2 データセットのサイズに関する事前実験

今回の分析では，クラスタ人気度の順位付けに写真の枚数を利用するため，単一の撮影者が同じ場所で同じ時間帯に連射した画像が多く含まれると，その画像が分析結果に影響を与えてしまう．そこで本研究では，分析前に全データセットの中から，連射などの影響を排除した部分データセットを取り出す．また，Mean Shift 法を適用するデータセットは小さいほど，計算時間が短くなるため，部分データセットのサイズは小さいほうがよい．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．そこで，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 1.9km 四方エリアとパリの 3.77km 四方エリア（両者の面積の差異は，Google Maps の仕様に依存している）を対象として，収集したデータの中から，30 分以内に同一撮影者によって撮影されたデータは 1 つとカウントした上で，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する．そして，それぞれのデータセットに対して bandwidth=0.001(100m) で Mean Shift 法を適用し，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．

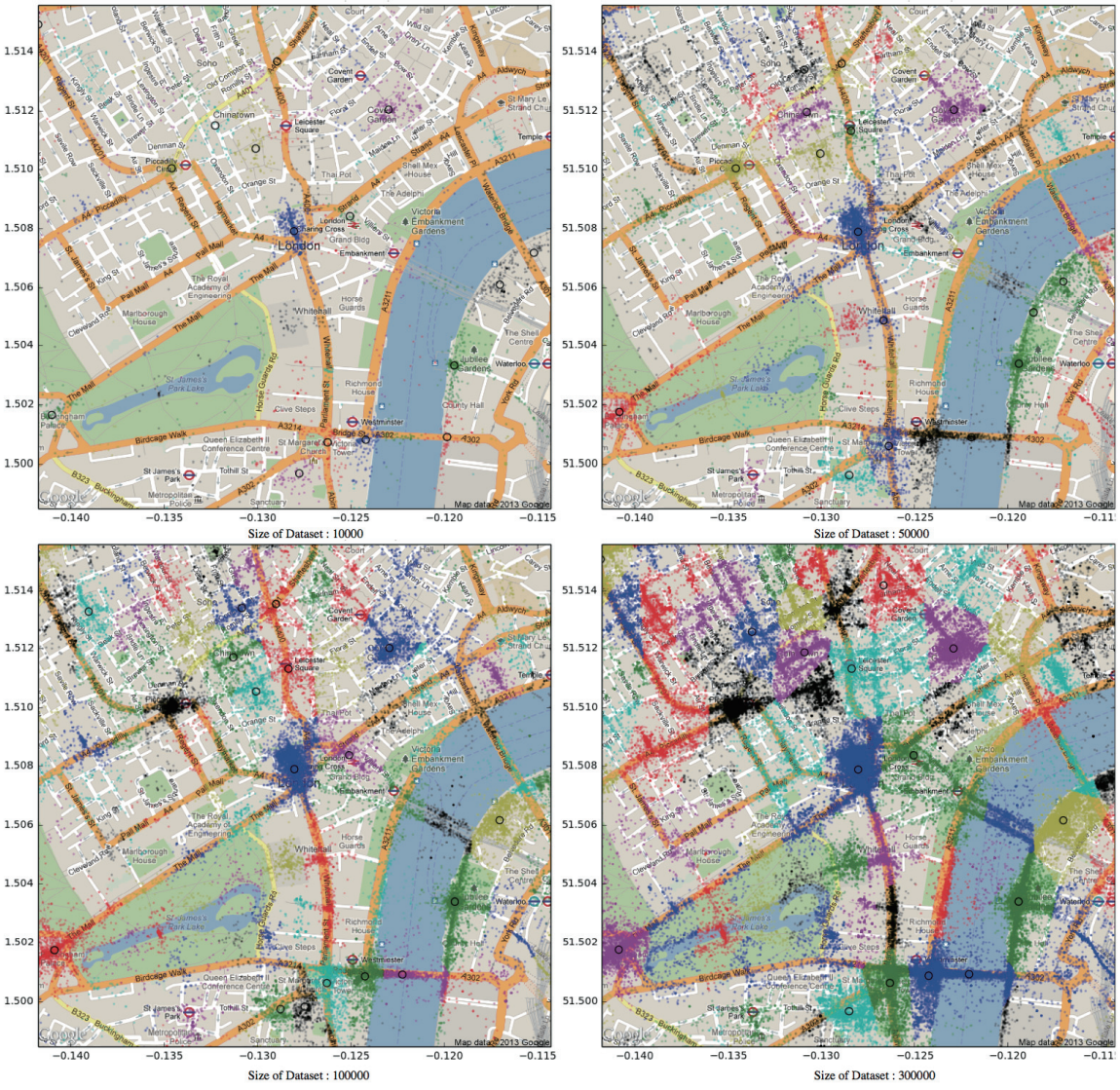

図 1　データセットサイズによるクラスタリング結果の違い（ロンドン）

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 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．

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

例えば,ロンドンの場合,1 辺は 1.9km であるため，10000 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．逆に，1 万枚のデータセット(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 3 から表 6 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わかる．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誤差はいずれも大きい．これはその POI の物理的な大きさが大きいために写真撮影地点(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と考えられる．

これらの結果を見ると，ロンドンに関しては，ランダムサンプリングによって得られた僅か 10000 件のデータセットでも,その 30 倍のデータセットと遜色ない結果が得られることがわかる．本誌には掲載していないが，より下位まで順位が必要な場合は，この限りではなく，より大きなデータセットが必要であると考えられる．

次に，ロンドンよりもデータセット当たりの面積を大きく設定したパリについて分析する．パリは，一辺が 3.77km であるため,各データセットの DBP は，昇順に，6.9，34.6，69.3，207.8 となる．同じ 10000 枚でも，ロンドンと比較して，DBP が極めて小さいため，見た目上はクラスタを認識しづらい（図は割愛）．しかしながら，個別に見ていくと，ロンドンと同様にこのような低 DBP（表７）であっても，高 DBP（表８）と遜色ない結果（ほぼ同等の 10 個の POI を選出し，その誤差も小さい）が得られている．順位に関しては，若干異なっていた（割愛したが，データセットが 50000 と 100000 の場合，１位 Louvre Pyramid，２位 Eiffel Tower となった）が，その誤差はどのデータセットでも同等（Eiffel Tower は約 11m，Louvre Pyramid は約 22m）であり，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結果であると考えられる．本研究では，枚数だけでなく，時間分散を加味した順位付けを提案しており，それによってこうした順位誤差も低減できるのではないかと考えている．

表３　Dataset Size = 10000 の結果（ロンドン）

	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	Trafalgar Square	51.508056	-0.128056	51.5079055	-0.128038862	16.79
2	The London Eye	51.5033	-0.1197	51.50332342	-0.11945936	16.91
3	British Museum	51.519459	-0.126931	51.51924954	-0.126861289	23.8
4	Tate Modern	51.507778	-0.099167	51.50788031	-0.09923166	12.24
5	Covent Garden	51.51197	-0.1228	51.51203413	-0.122969168	13.74
6	Piccadilly Circus	51.51	-0.134444	51.51003761	-0.134596073	11.36
7	Royal Festival Hall	51.505836	-0.116789	51.50608762	-0.117013619	32.05
8	Big Ben	51.500756	-0.124661	51.50079601	-0.1241917	32.89
9	Buckingham Palace	51.501	-0.142	51.50164185	-0.141012851	98.99
10	Parliament Square	51.500556	-0.126667	51.50071057	-0.12623318	34.69

表４　Dataset Size = 50000 の結果（ロンドン）

	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	Trafalgar Square	51.508056	-0.128056	51.50788021	-0.128021995	19.70
2	The London Eye	51.5033	-0.1197	51.50337771	-0.119408265	22.02
3	British Museum	51.519459	-0.126931	51.51926686	-0.126871105	21.78
4	Tate Modern	51.507778	-0.099167	51.50789364	-0.09923061	13.6
5	Covent Garden	51.51197	-0.1228	51.51202463	-0.122886121	8.53
6	Piccadilly Circus	51.51	-0.134444	51.51003115	-0.13457096	9.47
7	Big Ben	51.500756	-0.124661	51.50085145	-0.124220506	32.38
8	Parliament Square	51.500556	-0.126667	51.50057757	-0.126369023	20.83
9	Royal Festival Hall	51.505836	-0.116789	51.50619057	-0.117028952	42.82
10	Buckingham Palace	51.501	-0.142	51.50174603	-0.140816821	116.79

表５　Dataset Size = 100000 の結果（ロンドン）

	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	Trafalgar Square	51.508056	-0.128056	51.50788257	-0.128019371	19.46
2	The London Eye	51.5033	-0.1197	51.50336906	-0.119428653	20.35
3	Tate Modern	51.507778	-0.099167	51.5078909	-0.099228349	13.26
4	St Paul's Cathedral	51.513611	-0.098056	51.51379002	-0.099013597	69.4
5	British Museum	51.519459	-0.126931	51.51927515	-0.12688193	20.74
6	Royal Festival Hall	51.505836	-0.116789	51.50615413	-0.117029657	39.14
7	Piccadilly Circus	51.51	-0.134444	51.51003717	-0.134580905	10.37
8	Covent Garden	51.51197	-0.1228	51.51201826	-0.122928562	10.42
9	Big Ben	51.500756	-0.124661	51.50084041	-0.124234837	31.05
10	Buckingham Palace	51.501	-0.142	51.50173073	-0.14086723	113.12

表６　Dataset Size = 300000 の結果（ロンドン）

	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	Trafalgar Square	51.508056	-0.128056	51.50787524	-0.128015861	20.3
2	The London Eye	51.5033	-0.1197	51.50337514	-0.119424553	20.87
3	Tate Modern	51.507778	-0.099167	51.50787947	-0.099258178	12.94
4	British Museum	51.519459	-0.126931	51.51926718	-0.126877012	21.67
5	Covent Garden	51.51197	-0.1228	51.51202197	-0.12290653	9.39
6	Royal Festival Hall	51.505836	-0.116789	51.50615148	-0.117029839	38.88
7	Piccadilly Circus	51.51	-0.134444	51.51003304	-0.134571192	9.57
8	Big Ben	51.500756	-0.124661	51.50084761	-0.124237255	31.14
9	Parliament Square	51.500556	-0.126667	51.50060541	-0.126310114	25.38
10	St Paul's Cathedral	51.513611	-0.098056	51.51367288	-0.098282778	17.18

表７　Dataset Size = 10000 の結果（パリ）

	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	48.85836692	2.294373682	11.89
2	Louvre Pyramid	48.860854	2.335812	48.86104189	2.335897987	21.83
3	Notre Dame de Paris	48.853	2.3498	48.85316491	2.349367549	36.65
4	Arc de Triomphe	48.8738	2.295	48.87382757	2.294992512	3.11
5	Pompidou Centre	48.860653	2.352411	48.8605276	2.35211698	25.69
6	Basilique du Sacré-Cœur	48.886694	2.343	48.88626403	2.34302319	47.85
7	Place de l'Hôtel de Ville	48.856667	2.351389	48.85674765	2.351397153	8.99
8	Musée d'Orsay	48.86	2.327	48.85998727	2.326398714	44.14
9	Pont des Arts	48.858333	2.3375	48.85844818	2.337483121	12.87
10	Musée du Louvre	48.860339	2.337599	48.86045942	2.339874631	167.51

表８　Dataset Size = 300000 の結果（パリ）

	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	48.85836334	2.294393578	10.52
2	Louvre Pyramid	48.860854	2.335812	48.86104888	2.335899708	22.61
3	Notre Dame de Paris	48.853	2.3498	48.85315696	2.349368515	36.16
4	Arc de Triomphe	48.8738	2.295	48.87382847	2.294996218	3.18
5	Pompidou Centre	48.860653	2.352411	48.86053575	2.352157794	22.70
6	Basilique du Sacré-Cœur	48.886694	2.343	48.88630321	2.343055059	43.65
7	Place de l'Hôtel de Ville	48.856667	2.351389	48.85671831	2.351355457	6.21
8	Pont des Arts	48.858333	2.3375	48.85843634	2.337532974	11.74
9	Place de la Concorde	48.865556	2.321111	48.86556988	2.32113341	2.25
10	Musée d'Orsay	48.86	2.327	48.86002636	2.326379711	45.61

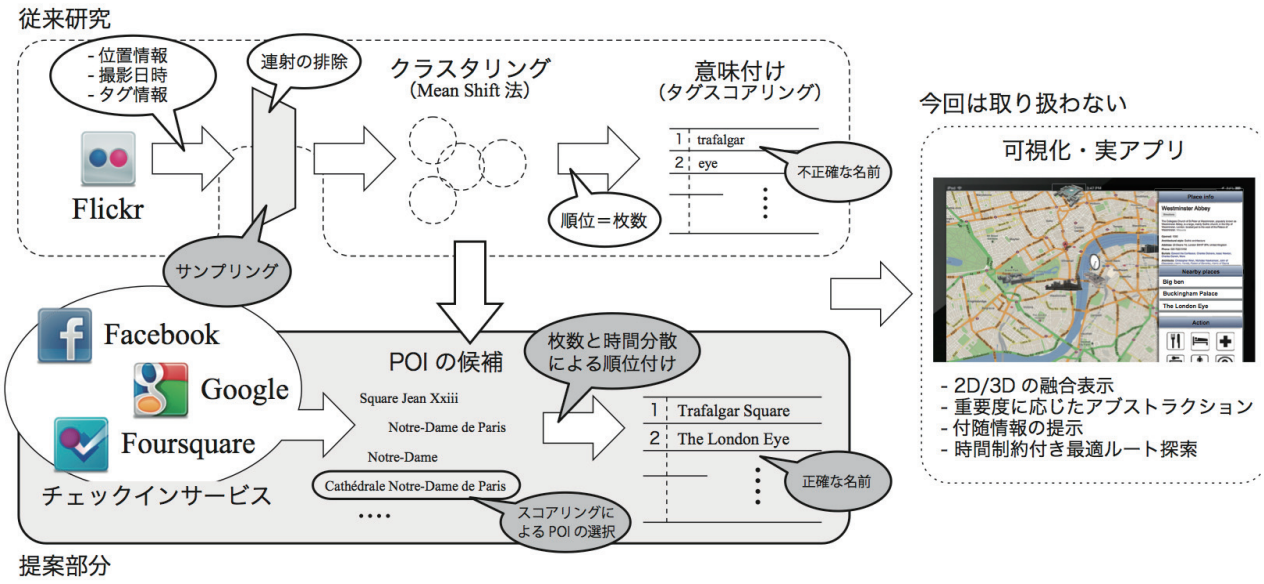

図 2　提案システムの構成

5 提案手法

本研究では，ソーシャルデータから実世界センシングを行う代表的な例として，旅行に主眼を置き，ソーシャル観光マップの自動生成を最終目的としている．今回の留学では，可視化システムに関しては取り扱わず，その前段階となるデータ分析部分に関して研究を進めた．図 2 はその構成と取り扱う部分について示したものである．データソースとして，Flickr 上の位置情報付き写真を利用し，Mean Shift 法でクラスタリングを行なって人気スポットを抽出するという全体の流れは，従来研究[1][3]と共通である．提案のコントリビーションは，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ランダムサンプリング手法，チェックインサービスからの情報を統合した POI 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．

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，撮影者当たりの写真の枚数は，28.4 枚になる．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，ひとりの撮影者が同じ場所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．さらに，提案では，古い写真を排除し(2004/01/01 00:00:00 以降の写真に限定し)，付与されている位置情報の精度が低い写真やタグが一切付与されていない写真も排除する．その結果，分析対象となる全データは，182 万枚に絞られる．この中からさらにランダムサンプリングを行う．今回，上位 10 位だけに焦点を当てるものとして，DBP が 20 以上となるデータセット（New York: 200000, San Francisco: 300000, London: 20000 , Paris: 50000, Berlin:100000）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，ロンドンは最もコンパクトにまとまっており，少ないデータ数でも DBP 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，旅行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 $\{s_1, s_2, \cdots, s_m\}$ の中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．

表９　ニューヨークにおける第５位クラスタに関するスコア計算の例

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

編集距離の計算は，有名な Levenshtein 距離でも良いが，今回は扱いやすさの観点（値が，文字列長にかかわらず，正規化された０～１で得られる）から Jaro-Winkler 距離［7］を利用する．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割ったもので POI 名 s をスコアリングする．単語数で除算する理由は，POI 名の長さの影響を減らすためである．また，ストップワード（the や of や記号など）は単語としてみなさず，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，出現順位で割ったものを POI 名 s_i $(i=\{1,2,\cdots,m\})$ のスコアとし，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理由は，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，文字列全体に見た時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件を候補として選択しているため，以降の評価では $m = 9$ となる．

表９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計算例を示す．人の目には，候補一覧からニューヨーク近代美術館と推測できるが，その表記はサービスによって異なっていることがわかる．この中で，最も他の候補との類似度が高いのは，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，The を除外するため１と数えることができ，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，提案手法によって別のソーシャルデータであるチェックインサービスから得た名前が適切である上，その表記もタグ分析の結果より優れていることがわかる．

5-3 時間分散を考慮した人気度推定に関して

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，わずか数枚の枚数差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．そこで，本研究では，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，そのスポットの旅行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いるとした時，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．この時，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 を用いて，クラスタ c に含まれる写真の撮影時間の分散 Dc を計算し，Dc にクラスタ内の写真の枚数を乗算した，Dc × k をクラスタ c の重要度と定義する．

4 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 件の比較を行った．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 10 から表 14 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わかる．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る．順位付けの評価は，今後，アプリケーションを市場に投入し，多人数に使ってもらう被験者実験を通じて行なって行きたいと考えている．

表 10　ロンドン

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

表 11　サンフランシスコ

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 12　ニューヨーク

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 13　パリ

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 14　ベルリン

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

【参考文献】

[1] Crandall, D., Backstrom, L., Huttenlocher, D. and Kleinberg, J.: Mapping the world's photos, *Proceedings of the 18th in- ternational conference on World wide web*, ACM, pp. 761– 770 (2009).

[2] Cheng, Y.: Mean shift, mode seeking, and clustering, Pat- tern Analysis and Machine Intelligence, IEEE Transactions on, Vol. 17, No. 8, pp. 790–799 (1995).

[3] Kurashima, T., Iwata, T., Irie, G. and Fujimura, K.: Travel route recommendation using geotags in photo sharing sites, *Proceedings of the 19th ACM international conference on In- formation and knowledge management*, pp. 579–588 (2010).

[4] Yin, Z., Cao, L., Han, J., Luo, J. and Huang, T.: Diversified trajectory pattern ranking in geo-tagged social media, Pro- ceedings of the Eleventh SIAM International Conference on Data Mining, SDM 2011, pp. 980–991 (2011).

[5] Kisilevich, S., Mansmann, F. and Keim, D.: P-DBSCAN: A density based clustering algorithm for exploration and analy- sis of attractive areas using collections of geo-tagged photos, *Proceedings of the 1st International Conference and Exhibi- tion on Computing for Geospatial Research & Application*, ACM, p. 38 (2010).

[6] Yang, Y., Gong, Z. et al.: Identifying points of interest by self-tuning clustering, *Proceedings of the 34th international ACM SIGIR conference on Research and development in In- formation*, ACM, pp. 883–892 (2011).

[7] 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).

〈発　表　資　料〉

題　　名	掲載誌・学会名等	発表年月
ソーシャルジオデータのクラスタリング結果に対する自動的な意味付けに関する一検討	情報処理学会全国大会	2013 年 3 月
Place API の統合	情報処理学会研究報告，モバイルコンピューティングとユビキタス通信研究会	2013 年 5 月