情報処理学会論文誌 Vol.55 No.9 2203–2215 (Sep. 2014)

推薦論文

知名度の地理的広がりを考慮した実世界スポットの地域局所性推定

徳永 陽子[1,a)] 数原 良彦[1] 佐藤 吉秀[1] 戸田 浩之[1] 鷲崎 誠司[2]

受付日 2013年12月17日, 採録日 2014年6月17日

概要：特定の住所を持つ実世界スポットについて，知名度の地域局所性推定に取り組む．実世界スポットには，全国的に知られているスポットや，周辺地域で局所的に知られているスポットなど，知名度の広がり方に違いがある．これらの実世界スポットを区別して提示することで，ユーザの検索満足度向上につながると考える．まず約 900 名の被験者に対して，全国 3 都市にある 150 件の実世界スポットについてアンケート調査を行った．被験者の居住地別に各実世界スポットを既知である人数比を求めた結果，実世界スポットの知名度の広がり方に差があることを確認した．本研究では，この知名度広がりを考慮した地域局所性を数値化したローカルスコアを定義する．また，ブログ記事において実世界スポット名と文書内共起している地名表現の地理的広さの違いを用いて，ローカルスコアを推定する手法を提案する．

キーワード：地域局所性，ローカルスコア，知名度広がり

Locality Inference of Real-world Spots Considering the Extent of Name Recognition

Yoko Tokunaga[1,a)] Yoshihiko Suhara[1] Yoshihide Sato[1] Hiroyuki Toda[1] Seiji Susaki[2]

Received: December 17, 2013, Accepted: June 17, 2014

Abstract: This paper shows how to infer attractive real-world spots known only to the locals. By emphasizing locally known spots over nationally famous spots, we can optimize users' satisfaction with search recommendations. First, 900 subjects undertook a questionnaire survey that listed 150 spots in Japan. The results, location of the subject and whether the spot was known, yielded the ratio of people who knew of the spots according to separation between their residence and the spots. We confirmed that the extent of name recognition is different for each spot. To permit automated data mining, we define "localization score" to measure the locality of each spot from user-entered text on the Web. Our achievement is to utilize place names that co-occur with the names of spots.

Keywords: Locality inference, Local score, Extent of name recognition

1. まえがき

近年，地域情報検索サービスの普及により，ユーザが外出前にパソコンやスマートフォンで訪問先の地域情報を調べる機会が増えている．ユーザが調べる地域情報の 1 つとして，外出の目的地に関する情報があげられる．本稿ではレストランや歴史的建造物など，特定の住所を持つ場所を実世界スポットと呼び，これを対象とする．

実世界スポットは，知名度の高さだけでなく，知名度の地理的な広がり方も様々である．たとえば，全国的によく

[1] 日本電信電話株式会社 NTT サービスエボリューション研究所
NTT Service Evolution Laboratories, NTT Corporation, Yokosuka, Kanagawa 239–0847, Japan
[2] 日本電信電話株式会社 NTT メディアインテリジェンス研究所
NTT Media Intelligence Laboratories, NTT Corporation, Yokosuka, Kanagawa 239–0847, Japan
[a)] tokunaga.youko@lab.ntt.co.jp

本稿の内容は 2013 年 9 月の FIT2013 第 12 回情報科学技術フォーラムにて報告され，同プログラム委員長により情報処理学会論文誌ジャーナルへの掲載が推薦された論文である．

知られている有名なものや，遠くの人には知られていないが地元ではよく知られているものなどがある．本稿では，任意の実世界スポットが所在する周辺の人だけに知られているという度合いを地域局所性と呼ぶ．

各実世界スポットの知名度の地理的広がり方を考慮し，地域局所性の高い実世界スポットを提示することができれば，ユーザが新たな目的地を発見することができる．たとえば，地元で人気のある食堂や桜が綺麗な公園などは，地域局所性が高い実世界スポットであり，有名な実世界スポットとは異なる穴場であることから，新たな発見となる可能性がある．また，地域局所性の推定が可能になれば，全国的に知名度が高い実世界スポットと，地元の人だけに知られている実世界スポットの分別ができるようになり，地域に関する知識やニーズなど，ユーザに合わせた情報提供が可能になると考えた．

従来のサービスでは，地域局所性という観点を利用して実世界スポットを提示することは困難であった．実世界スポット情報を提供する既存サービスとして，来場者数や口コミの量・レビュー点数などを用いて人気の実世界スポットをユーザに提示するものがある[*1,*2]．なじみのない場所に出かけるユーザには，行き先周辺に関する知識が少なく，このようなサービスによって得られた観光客向けの実世界スポット情報は外出時の行動支援として有用であると考えられる．しかし，ユーザが居住地域周辺で訪問先を決定する場合や，何度も訪問した地域へ訪れる場合，周辺の有名な実世界スポットはすでに把握している可能性が高い．このような場合，ユーザが求める情報とは異なるため，提示する必要性は低いと考えられる．これまでは，この地域局所性を測る尺度が存在しなかったため，知名度の広がり方による実世界スポットの区別はできなかった．

本研究では，まず実世界スポットの知名度広がりを考慮した地域局所性の定量的な尺度を定める．人手評価データを用いて実世界スポットの知名度を特定し，それに基づいて地域局所性を計算する手法を定義する．

また，任意の実世界スポットについて地域局所性を求める場合，人手評価データの作成は困難であるため，地域局所性を自動推定する必要がある．そこで，実世界スポットを訪問した記述が多く含まれるブログ記事を情報源として利用し，スポット名と共起する地名を用いて，任意の実世界スポットの地域局所性を推定する手法を提案する．

本研究では，(1) 実世界スポットの地理範囲別の知名度を特定し，その知名度を用いて (2) 実世界スポットの地域局所性を求めるという流れで取り組んだ．本稿の流れを**図 1** に示す．2 章では，人手で作成した評価データを (1) 地理範囲別の知名度として用いた際の (2) 地域局所性の計算方法について定義する．3 章では，任意の実世界スポットに対し，ブログ記事を用いて (1) 地理範囲別の知名度を推定する方法を提案する．4 章では，前章で得た知名度の推定値と，それを用いて (2) 地域局所性を計算した結果の評価を行い，その分析結果について報告する．

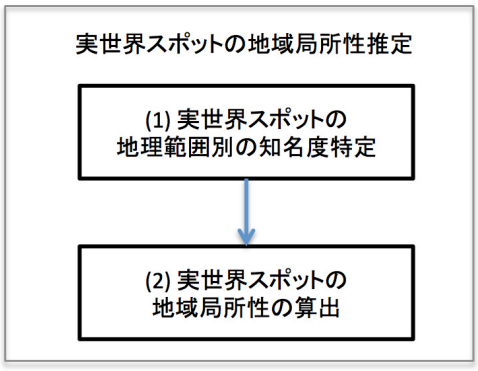

図 1 本稿の流れ
Fig. 1 The story of this paper.

2. 地域局所性の数値化

この章では，実世界スポットの地域局所性を数値化する方法について示す．まず人手評価データについて述べた後，その分析結果に基づいて地域局所性を数値化する手法を定義する．

2.1 評価データの作成

本実験では，知名度の広がり方を定式化するために，ある実世界スポットを知っている人の割合が，スポット周辺から範囲が広がるにつれてどのように変化するのかを調べた．実験で用いたエリアを**表 1** に示す．人口が集中する都心を含む関東エリア，観光地が多い関西エリア，県外からの人の流入が比較的少ない九州エリアの 3 つを拠点として選択し，各エリアについて代表 3 都市（横浜市，京都市，福岡市）を「拠点市」，拠点市がある府県を「拠点県」，拠点県に隣接する都道府県を「隣接県」，隣接県を除く他の都道府県を「他県」と呼ぶ．

調査は，拠点市内の住所を持つ実世界スポットを対象とし，被験者に対してそのスポットを知っているかどうか質問するアンケート形式とした．被験者には「“横浜市” にある “日産スタジアム” を知っていますか」という質問形式でスポット名と拠点市を提示し，
(1) 名前も場所も知っている，
(2) 名前は知っているが場所は知らない，
(3) 知らない，
の 3 つの選択肢の中から選んで回答してもらった．

対象となる実世界スポットは，ウェブから独自にクロールした文書からスポット名と住所の組を抽出し，横浜市から 53 個，京都市から 57 個，福岡市から 40 個，計 150 個を選定した．

*1 http://www.rurubu.com/
*2 http://www.mapple.net/

表 1 評価に用いたエリアとその区域
Table 1 Areas and their regions used for the evaluation.

エリア	区域			
	拠点市	拠点県	隣接県	他県
関東	横浜市	神奈川県	東京都 千葉県 静岡県 山梨県	京都府・福岡県とその隣接県
関西	京都市	京都府	大阪府 滋賀県 奈良県 三重県 兵庫県 福井県	神奈川県・福岡県とその隣接県
九州	福岡市	福岡県	山口県 大分県 長崎県 熊本県 佐賀県	神奈川県・京都府とその隣接県

表 2 エリアと区域ごとの被験者数（単位：人）
Table 2 The number of annotators for each region in each area (unit: people).

エリア	区域			
	市内	県内市外	隣接県	他県
関東	58	54	165	629
関西	68	45	228	515
九州	67	46	175	618

知名度の広がり方を求めるため，居住地に基づいて被験者を選定した．被験者は，現在の居住地に基づいて各エリアの拠点市内（以下，市内とする），拠点県内の拠点市以外の市町村（以下，県内市外とする），隣接県，他県の 4 つの範囲（以下，区域とする）に分けて選んだ．このとき，たとえば隣接県に在住の被験者は過去に拠点市や拠点県に住んだことがなく，通勤・通学をしたこともないことを条件にするなど，過去の居住歴や通勤通学歴なども考慮した．エリア別・区域別の被験者数は**表 2** に示す．

2.2 各エリアの特徴

各エリアの拠点県について，観光庁の「観光入込客統計」[*3]および拠点市について「観光入込客数及び観光消費額」[*4,*5]から，平成 22 年の観光入込客数を**表 3** に示す．3 つのエリアのうち，関西エリアの京都府では，拠点県への観光入込客のうち県外からの観光入込客の割合が 76%と最も多い．さらに，県外から拠点市である京都市への観光入込客も，全体の 58%を占めている．一方，九州エリアでは，拠点県の福岡県への観光入込客のうち，県外からの観光入込客の割合は 33%しかおらず，県外から拠点市の福岡市への観光入込客も 10%未満であった．このことから，関西エリアで拠点県の京都府および拠点市の京都市は，府外からの観光客が多く訪れるという特徴があることが分かる．また，九州エリアで拠点県の福岡県および拠点市の福岡市は，県内での人の流れが多く，県外からの流入が少ないという特徴があることが分かった．関東エリアの拠点県の神奈川県は，県内外ともに観光客が多いという特徴があると考えられる．

*3 http://www.mlit.go.jp/kankocho/siryou/toukei/irikomi.html
*4 http://www.pref.kyoto.jp/kanko/1317298591939.html
*5 http://www.city.fukuoka.lg.jp/keizai/shukyaku/shisei/001_2_2_2.html

2.3 評価データの分析

評価データをもとに，実世界スポットの知名度について分析を行った．ここでは，質問に対して「(1) 名前も場所も知っている」を選んだ被験者のみを，その実世界スポットを知っている適合者として扱う．それ以外の回答を選んだ被験者は，その実世界スポットを知らないとする．各スポットについて，被験者全体のうちの適合者の比で降順に並べたグラフを**図 2** に示す．この図から，選択した 150 個の実世界スポットには，多くの人に知られている実世界スポットと少数の人に知られている実世界スポットが混在していることが分かる．

次に，区域別の知名度の違いを見る．ここでは，ある実世界スポットについて，区域別の適合者比を，全区域の適合者比の合計が 1 になるように正規化した値をその区域での知名度とする．すなわち，区域ごとに知名度が定義される．この際，隣接県と他県の知名度の和を県外の知名度とする．区域別に見た場合の知名度の違いの代表例として，京都市内の 3 つの実世界スポットの知名度を**図 3** に示す．一般に，実世界スポットが存在する場所に近い範囲では知名度が高く，範囲が広くなり実世界スポットから遠くなるほど知名度が低くなる．図 3 にあげた 3 つの実世界スポットも，市内での知名度が最も多く，県内市外，県外と範囲を広げるにつれて知名度が少なくなっている．

さらに，実世界スポットによって知名度の減衰の度合いが異なっており，知名度の広がり方にも違いがある．図 3 の大石神社は，市内だけで顕著に知名度が高く，それ以外では低くなっていることから，京都市以外の人には同じ府内であっても知られておらず，拠点市内でのみ知名度が広がっていることを示している．また，図 3 の東寺は，全区域で知名度の差が小さいことから，東寺の近くだけでなく広い範囲でよく知られており，全国的に知名度が広がっていることを示している．本稿では，上記で例にあげた大石神社に見られるような，知名度がその実世界スポット周辺に偏っていることを地域局所性が高いと定義する．

また，真珠庵と大石神社を比較すると，大石神社の方が地域局所性が高い．真珠庵は市内と県内市外の知名度の差が少ないが，県外まで範囲が広がると知名度が急に低くなっている．これは，京都府内の人までは知られているが，府外の人にはあまり知られておらず，拠点県内で知名度が広がっていることを示している．このことから，県内市外

表 3 観光客数と県外からの入込数．() 内は拠点県の観光客数に対する割合
Table 3 The numbers of tourist. Percentage of tourist for the prefecture is in ().

エリア（拠点県・市）	拠点県の観光客数	県外から拠点県への観光客数	県外から拠点市への観光客数
関東（神奈川県横浜市）	103,201	44,350（43%）	該当データなし
関西（京都府京都市）	76,741	58,600（76%）	44,349（58%）
九州（福岡県福岡市）	100,126	32,652（33%）	9,690（10%）

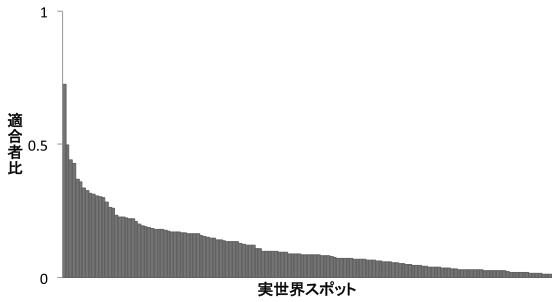

図 2 各実世界スポットの適合者比
Fig. 2 The ratio of relevant annotators for each real-world spot.

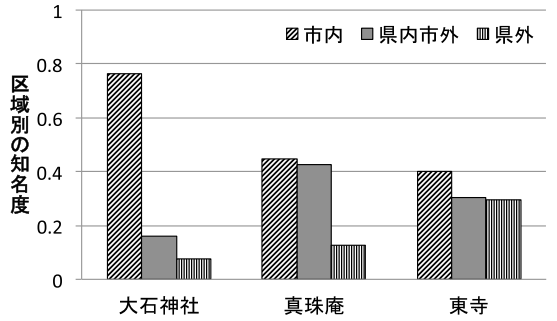

図 3 各実世界スポットの区域別知名度
Fig. 3 The name recognition of real-world spots in each region.

と県外の知名度の差よりも，市内と県内市外との差が大きい方がより地域局所性が高いことが分かる．

このように，地域局所性は図 2 に示した実世界スポットの適合者比では表現できない特徴を表すことができる．東寺のように区域別の知名度の差が小さいスポットより，大石神社のように差が大きいスポットのほうが地域局所性が高いことから，地域局所性の大きさは区域別の知名度の差による影響が大きいと考えられる．また，真珠庵のように県内市外と県外の差が大きいスポットより，大石神社のように市内と県外市外の差が大きいスポットのほうが，より地域局所性を高くすることが望ましい．次の節では，この地域局所性の大きさを数値化する方式について検討する．

2.4 ローカルスコアの定義

評価データの分析結果をふまえて，地域局所性を数値化する方式について検討する．ここでは，この地域局所性を数値化したものをローカルスコアと呼ぶこととする．

前述のとおり，知名度の広がりの違いが地域局所性に影響するため，スポット s の市内の知名度を RT_s，県内市外の知名度を RC_s，他県の知名度を RP_s とし，知名度の差を用いてローカルスコア LS を次の式で定義する．

$$LS(s) = \lambda(RT_s - RC_s) + (1-\lambda)(RC_s - RP_s) \quad (1)$$

ただし，λ は $0 \leq \lambda \leq 1$ の定数とする．この式では，$(RT_s - RC_s)$ が市内と県内市外の知名度の差，$(RC_s - RP_s)$ が県内市外と県外の知名度の差を表している．これによって，知名度の広がり方を考慮して，地域局所性が大きいほど値が大きくなるようなローカルスコアとして数値化できる．

実際に，$\lambda = 0.75$ として，150 個のスポットのうち，市内で 10%以上の人が知っていると答えたスポット 131 件（横浜市で 49 件，京都市で 43 件，福岡市で 39 件）について，エリア別にローカルスコアが大きい順にランキングをした．ランキングの上位 5 件下位 5 件を**表 4** に示す．ランキングの上位には，地元の人が訪れる公園など，知名度が局所的なスポット，ランキング下位には有名なお寺や大規模なコンサートホールなど，知名度が全国的に広がっているスポットが並んでいる．このように，定義したローカルスコアによって，各スポットの知名度の広がりを反映した地域局所性を数値で表現することができるようになった．

本研究で提案するローカルスコアとは，ある実世界スポットの知名度が，実世界スポットから遠ざかるにつれてどのように変化するかを表す値であり，知名度の絶対的な高さを表すものではないことに注意されたい．つまり，ローカルスコアは，知名度の高さにかかわらず，その広がり方のみに注目し，局所的な広がりなのか，全国的な広がりなのかを数値で示すものである．今後，知名度の高い実世界スポットの地域局所性を知りたい，といった状況に対応するためには，「実世界スポットを知っている人数」の観点を取り入れる必要があるが，この観点は今回提案したローカルスコアでは扱わないものとする．我々は実世界スポットの知名度の広がりは知名度の絶対的な値によらず定めることができると考えたため，本稿で定義するローカルスコアの対象外とした．なお，実世界スポットの知名度の高さは，たとえば実世界スポット名を含むブログ記事の件数などで推定可能であると考えており，これらを組み合わせることでローカルスコアを拡張できると考えている．

3. 地名共起を用いたローカルスコア推定

前章では，人手で評価した知名度のデータを用いてローカルスコアについて定義した．しかし，すべての実世界ス

表 4 ローカルスコアによるランキング
Table 4 Ranking by local score.

ローカルスコア	関東	関西	九州
最大	掃部山公園	大石神社	聖福寺
	田谷の洞窟	加茂別雷神社	パピオアイスアリーナ
	横浜メディアビジネスセンター	勧修寺	東長寺
	MotionBlue 横浜	大原野神社	雁の巣レクリエーションセンター
	都筑中央公園	大河内山荘	山王公園
	…	…	…
	日産スタジアム	二条城	マリンメッセ福岡
	新横浜公園	西雲院	福岡タワー
	港の見える丘公園	清水寺	海の中道海浜公園
	美しが丘公園	正法寺	博多バスターミナル
最小	横浜アリーナ	弘源寺	キャナルシティ博多

ポットについて人手によるデータを利用することはできないため，人手を用いずにローカルスコアの推定値を計算する必要がある．そこで，本研究では実世界スポットについて述べる際に用いる地名の地理的広さに着目してローカルスコアの推定を試みる．

3.1 予備実験

スポット名と共起している地名と言及している人の居住地の関係を確かめるために，次の実験を行った．マイクロブログサービス Twitter[*6]では，投稿者の居住地をプロファイルに登録することができる．そこで，この居住地の情報を用いて，実世界スポットの所在都道府県との一致と用いる地名の関係を調べた．

対象の実世界スポットは独自に収集した有名な実世界スポットから無作為に選択した 9 個とする．プロファイルに投稿者の居住都道府県が記載されているツイート記事のみを対象とし，9 個のうちいずれかの実世界スポット名と，その実世界スポットの住所を包含する地名の両方を含む記事，計 3,229 件を用いた．利用した実世界スポットと，それぞれの実世界スポット名を含むツイート件数を**表 5** に示す．

すべての記事に対して，投稿者の居住都道府県のカテゴリ別に，言及している地名の地理的広さ別の頻度を数えた．まず，対象の記事に含まれている地名について，その地名の地理的広さを判別する．地名の地理的広さは，市町村よりも細かい地名を町レベル，市町村を表す地名を市レベル，都道府県を表す地名を県レベルの 3 つのレベルとする．次に，その記事の投稿者の居住都道府県が，記事内で言及されている実世界スポットが存在する都道府県と同一か，隣接した都道府県か，その他の都道府県かの 3 つのカテゴリに分別する．

結果を**図 4** に示す．投稿者が対象の実世界スポットと同一都道府県に住んでいる場合，町レベルでスポット名を用いて言及することが多い．一方，隣接都道府県やその他の都道府県など，投稿者の住んでいる場所が対象のスポットから遠くなるにつれて，市レベルや県レベルを用いた言及が多くなっている．この結果から，実世界スポットについて言及する際には，投稿者が住んでいる場所と実世界スポットが近い場合は地理的に狭い範囲を表す地名，遠い場合には地理的に広い範囲を表す地名を用いるというように，用いる地名のレベルが異なる傾向があることが示唆さ

[*6] http://twitter.com

表 5 検証に用いた実世界スポット名称と各実世界スポット名称を含むツイート件数
Table 5 The real-world spots used for the verification and numbers of tweets that contain the spot's name.

実世界スポット名	所在都道府県	ツイート件数
スカイツリー	東京都	2,151
ディズニーランド	千葉県	326
関帝廟	神奈川県	5
ランドマークタワー	神奈川県	55
三溪園	神奈川県	18
コスモワールド	神奈川県	12
清水寺	京都府	291
平安神宮	京都府	87
大宰府	福岡県	284
合計		3,229

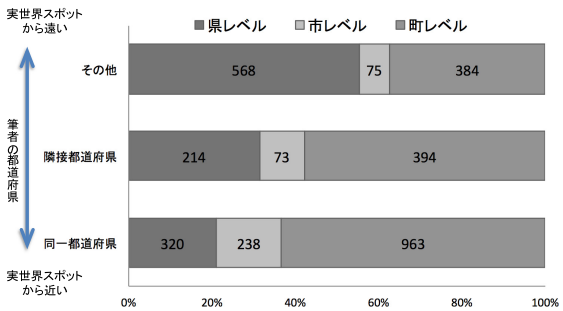

図 4 ツイート投稿者の居住区域と共起地名の関係
Fig. 4 Relation between the region of a Tweet author's residence and a co-occurring topological word.

れた．

この予備実験では，居住地の情報が必要となるため，記事数が多い有名な実世界スポットに限定してツイート記事を用いたが，ローカルスコアの推定ではツイート記事は情報源として適さないと考えられる．実際のツイート記事の中には，地名の代わりに現在地の緯度経度の情報を添付して投稿されているものも多く見られる．しかし，緯度経度の情報では地名の地理的広さが得られないため，今回の検証では対象外とした．このようなツイート記事の特徴から，言及されている記事数が少ない実世界スポットでは対象となる記事数を十分に確保できない可能性がある．

3.2 ローカルスコアの推定

予備実験の結果，実世界スポットの周辺に住む人と広い区域に住む人では，言及する際に用いる地名のレベルが異なる傾向が分かった．これは，実世界スポットに近い人はその周辺に詳しいため，詳細な地名を知っている一方で，遠い人はその周辺に関する知識が少なく，詳細な地名を知らないことが原因の 1 つであると考えられる．たとえば横浜市内に住んでいる人は，横浜市内の実世界スポットについて言及する際に，「神奈川県」や「横浜市」という地名は自明であるため，それより細かい「伊勢佐木町」「石川町」などの横浜市内のどの地域かを特定するために必要な細かい地名を用いる．同様に，神奈川県内の横浜市以外に住んでいる人は，横浜市内の実世界スポットについて言及する際に，「神奈川県」という自明な地名は用いないが，県内のどの地域かを特定するために市名を用いる．さらに，神奈川県外に住んでいる人は，横浜市内の実世界スポットについて言及する際に，まず日本国内のどの地域かを特定するために県名を用いる．

この傾向から，あるスポットについて言及する際に用いる地名が

- 町レベルの場合：市内の人が言及している
- 市レベルの場合：県内市外の人が言及している
- 県レベルの場合：県外の人が言及している

と仮定することで，言及した人が住んでいる区域を推定できる．これをもとに，式 (1) で用いた知名度 RT_s，RC_s，RP_s の推定を行うことで，人手による評価データがない場合でも実世界スポットのローカルスコアを推定できると考えた．この知名度の推定値を推定知名度と呼ぶ．

本稿では，ブログ記事を用いて推定知名度を算出する．ブログ記事は，つねに決まったレベルの地名で記述するニュース記事などとは異なり，投稿者の感覚にあったレベルの地名で記述されていると考えられる．そこで，ブログ記事を情報源として，実世界スポットの区域別推定知名度を求める．

推定知名度の具体的な計算方法について述べる．まず，対象とする実世界スポットのスポット名を含むブログ記事のうち，実世界スポットの場所を表す地名を含む記事のみを対象とする．次に，記事内に含まれている地名が町レベル・市レベル・県レベルのどれに該当するか判定し，レベル別に記事数を数える．全レベルでの合計記事数が 1 になるように，各レベルの記事数を正規化した値を，対象の実世界スポットを知っている適合者比の推定値，つまり推定知名度として用いる．

ある実世界スポットのスポット名を s，町レベルの地名を geo_{town}，市レベルの地名を geo_{city}，県レベルの地名を geo_{pref} とする．スポット名 s といずれかのレベルの地名を含むブログ記事数を D_s，任意のレベルの地名 geo_x を含むブログ記事数を $d(s, geo_x)$ とし，市内の推定知名度を RT'_s，県内市外の推定知名度を RC'_s，他県の推定知名度を RP'_s としたとき，下記のように算出する：

$$RT'_s = \frac{d(s, geo_{town})}{D_s},$$

$$RC'_s = \frac{d(s, geo_{city})}{D_s},$$

$$RP'_s = \frac{d(s, geo_{pref})}{D_s}.$$

ただし，

$$D_s = \sum_{x \in \{town, city, pref\}} d(s, geo_x)$$

とする．この推定知名度を用いて，推定ローカルスコアを求めることができる．推定ローカルスコア LS' を次の式で定義する．

$$LS'(s) = \omega(RT'_s - RC'_s) + (1 - \omega)(RC'_s - RP'_s) \quad (2)$$

ただし，ω は $0 \leq \omega \leq 1$ の定数とする．

4. 評価実験

この章では，提案知名度が人手で作成した評価データに基づく知名度を正しく近似しているかを確かめるための評価実験とその結果について述べる．

4.1 データセット

実験に用いたデータについて述べる．対象とした実世界スポットは，前章の評価データを作成する際に用いた実世界スポットのうち，スポット名を含むブログが 10 件以上存在するスポット計 131 個とした．ブログ記事は，日本国内の大手ブログサイト 10 件以上から，記事を投稿した際に得られる RSS をもとに独自に収集したものを利用し，2012 年 1 月から 2013 年 3 月までの日本語ブログ記事の中から，対象とする実世界スポット名を含む記事のみを用いた．前処理として，1 つのブログ記事中に 10 個以上の異なる地名を含む記事は，スパムやアフィリエイトの可能性が高いと見なし，除外した．

4.2 実験条件

ブログ記事を用いた推定知名度の計算方法について述べる．まず，ブログ記事中で対象とする実世界スポット名と，そのスポットが存在する住所を包含する地名が共起するかどうかを解析した．地名の抽出には，記事中に出現した地名表現について，周辺に出現する語や地名の有名度などを手がかりに正しい地名を特定する手法 [1] を用いた．このとき，地名は後方一致のみを見ることとし，たとえば「京都府京都市左京区岡崎西天王町」の場合は，岡崎西天王町という町レベルの地名が書かれているものとして扱った．また，関西エリアと九州エリアについては，京都府京都市，福岡県福岡市というように都道府県名と市名が同じである．そのため，記事中に「京都」または「福岡」とのみ書かれた場合には，同一記事内に書かれている内容を考慮しても，県レベルと市レベルの判定が不可能な場合もあると考えられる．そこで，本実験では市レベルと町レベルの中間である区レベルの地名まで記事中に書かれている場合，市レベルの住所が書かれていると判断した．

また，1 つの記事中に複数の地名が含まれている場合，最も詳細なレベルの地名のみを選択して扱った．次に，比較した推定ローカルスコアについて述べる．本研究では，推定知名度を利用しない 3 手法，推定知名度の差を用いる 7 手法，推定知名度をそのまま用いる 1 手法の計 11 手法を比較した．

LS'_IDF：実世界スポット名を含むブログ記事数の逆数．

LS'_GEOIDF：実世界スポット名と，その所在地を包含する地名を 1 つ以上含むブログ記事数の逆数．

$LS'\_0.1$：式 (2) で，$\omega = 0.1$ としたもの．

$LS'\_0.25$：式 (2) で，$\omega = 0.25$ としたもの．

$LS'\_0.75$：式 (2) で，$\omega = 0.75$ としたもの．

$LS'\_0.9$：式 (2) で，$\omega = 0.9$ としたもの．

LS'_TC：式 (2) で，$\omega = 1.0$ とし，$(RT'_s - RC'_s)$ を推定ローカルスコアとしたもの．

LS'_TP：式 (2) で，$\omega = 0.5$ とし，$(RT'_s - RP'_s)$ を推定ローカルスコアとしたもの．

LS'_T：$(RT'_s - (RC'_s + RP'_s))$ を推定ローカルスコアとしたもの．

LS'_RT：実世界スポットの市内の推定知名度を推定ローカルスコアとしたもの．

LS'_DF：実世界スポット名を含むブログ記事数．

推定知名度を利用せずに地域局所性を表す 2 手法について述べる．LS'_IDF は，スポット名を含む文書数が多いほど，その実世界スポットはよく知られているスポットであると考えられるため，その逆数をとることで全国的にはあまり知られていないスポットが上位になると考えられる．LS'_GEOIDF は，地名のレベルを問わず，スポット名と地名が共起している文書数の逆数である．ブログ記事内に実世界スポットと地名が共起している場合，その実世界スポットに行った経験などを記述している場合が考えられる．よって，投稿者が対象の実世界スポットについて知っている可能性が，地名が含まれていない記事よりも高く，地名共起の有無にかかわらずスポット名を含む全文書数を用いた LS'_IDF よりも正しくローカルスコアを推定できると予想される．

続いて，推定知名度の差を用いた 7 手法について述べる．$LS'\_0.1$，$LS'\_0.25$，$LS'\_0.75$，$LS'\_0.9$ は式 (2) の ω の値を変えたもので，ω の値が大きくなるほど市内と県内市外の知名度の差による影響が大きくなる．LS'_TC と LS'_TP は，同様に式 (2) の ω の値を 1.0 または 0.5 として計算した値である．ω の値を 1.0 として計算した LS'_TC は，市レベルと町レベルの推定知名度の差のみを用いて推定ローカルスコアとする．同様に，ω の値を 0.5 として計算した LS'_TP は，県レベルと町レベルの推定知名度の差のみを用いて推定ローカルスコアとする．この 2 手法と他の手法を比較することで，どの区域間の知名度の差が実際のローカルスコアと関係が深いかを調べることができる．また，LS'_TP は市レベルの地名を用いずに推定知名度を求めている．LS'_T は，県内市外と県外を “市外” という一括りの区域と見なし，市レベルと県レベルの推定知名度の和を市外の知名度とし，町レベルの推定知名度との差を推定ローカルスコアとした．これによって，市内・県内市外・県外という区切り方ではなく，市内・市外という区切り方で知名度を測り，その差を用いることでローカルスコアを適切に推定できるかどうかを調べる．

LS'_RT は，推定知名度の差ではなく，町レベルの推定知名度をそのまま推定ローカルスコアとして用いたものである．これによって，実世界スポットを包含する最も小さい区域での知名度と，知名度の広がりを表すローカルスコアとの関係を調べることができる．LS'_DF は，LS'_IDF と同様にスポット名を含む文書数を用いるが，逆数をとらず文書数そのものを推定ローカルスコアとして用いる．つまり，実世界スポットに関する文書が多いほどスコアが高くなるため，知名度の広がりを表す手法とは異なる指標であるととらえられる．

今回の実験では，京都や福岡のように県レベルと市レベルが同じ地名の場合を考慮し，市レベルについては区レベルの地名まで書かれた場合のみをカウントすることで曖昧性を回避している．しかし，県レベルの地名については．明確に「京都府」「福岡県」と書かれている場合と，「京都」「福岡」としか書かれておらず，地名を抽出する際に県レベルか市レベルかを記事中の情報から推定した結果，県レベルと判定された場合も数に含まれている．この場合，地名の抽出手法の精度によって推定知名度の精度が影響を受ける可能性があると考えられる．そこで，曖昧性のない町レベルの地名と，地名の抽出手法の精度によって影響を受ける県レベルの地名のみを用いて算出した LS'_TP による推

表 6 ランキング比較評価
Table 6 Comparative evaluation of ranking methods.

(a) LS_0.75 を正解とした場合

	P@15			nDCG		
	横浜	京都	福岡	横浜	京都	福岡
LS'_IDF	0.47	**0.60**	0.73	0.78	0.89	**0.98**
LS'_$GEOIDF$	0.53	0.53	0.73	0.83	0.87	0.92
LS'_0.1	0.33	0.53	0.80	**0.90**	0.89	0.97
LS'_0.25	0.33	0.53	**0.87**	**0.90**	0.89	0.96
LS'_TP	**0.60**	0.53	0.80	0.80	0.90	0.95
LS'_0.75	**0.60**	0.53	0.73	0.80	0.90	0.94
LS'_0.9	**0.60**	0.53	0.73	0.80	**0.91**	0.94
LS'_TC	0.33	0.20	0.20	0.79	0.70	0.72
LS'_T	**0.60**	0.53	0.73	0.80	0.90	0.94
LS'_RT	**0.60**	0.53	0.73	0.80	0.90	0.94
LS'_DF	0.07	0.13	0.07	0.72	0.68	0.68

LS_0.25 を正解とした場合

	P@15			nDCG		
	横浜	京都	福岡	横浜	京都	福岡
LS'_IDF	0.20	**0.67**	**0.60**	0.82	**0.90**	**0.90**
LS'_$GEOIDF$	0.20	**0.67**	**0.60**	0.74	**0.90**	**0.90**
LS'_0.1	0.20	0.53	0.40	0.79	**0.90**	0.85
LS'_0.25	0.27	0.53	0.47	0.79	**0.90**	0.80
LS'_TP	0.40	0.53	0.40	0.90	**0.90**	0.76
LS'_0.75	0.40	0.53	0.40	**0.91**	**0.90**	0.76
LS'_0.9	0.40	0.53	0.40	0.90	**0.90**	0.76
LS'_TC	**0.53**	0.07	0.20	0.83	0.68	0.77
LS'_T	0.40	0.53	0.40	**0.91**	**0.90**	0.76
LS'_RT	0.40	0.53	0.40	**0.91**	**0.90**	0.76
LS'_DF	0.33	0.07	0.20	0.77	0.67	0.70

定知名度を調べることで，地名の抽出手法による影響を調べることができると考えた．

実験に用いた評価指標について述べる．エリアをクエリとし，各実世界スポットを 1 文書と見なすと，各エリアについて，実世界スポットのランキング問題であると考えることができる．そこで，情報検索分野でランキング評価に用いられる指標を用いて，推定ローカルスコアの評価を行った．評価指標は，

- 各エリアの LS によるランキング上位 1/3 を正解としたときの適合率 P@15,
- 各エリアの LS によるランキング上位 1/3 を 3 点，下位 1/3 を 1 点，残りを 2 点としたときの nDCG [2],

の 2 つを用いた．また，正解ローカルスコア LS として，以下の 2 種類を用いた．

- LS_0.75：$\lambda = 0.75$ とした場合の LS
- LS_0.25：$\lambda = 0.25$ とした場合の LS

2 つの評価指標の解釈について述べる．P@15 は，15 位以内に含まれる正解の割合を表した指標である．したがって，P@15 の値が大きいほど，正解とした上位 1/3 のスポットについて，順位によらず上位 15 個以内により多くランキングすることができたと解釈できる．nDCG とは，適合文書の適合度合いを点数に置き換えて，検索順位の上位にある文書に重みをかけた指標である．よって，nDCG の値が大きいほど，当該手法によって LS が高いスポットについて適切に上位にランキングできたことを示す．

4.3 評価結果と考察

評価の結果を**表 6** に示す．各評価値についてエリア別に見たとき，最も大きい値を太字で示した．まず，LS_0.75 を正解とした場合と LS_0.25 を正解とした場合を比較する．特に P@15 で比較した場合，LS_0.25 を正解とした結果は，LS_0.75 を正解とした結果に比べ，LS'_DF を除き，評価値が下がったものが多かった．このことから，今回提案した推定知名度に基づく推定ローカルスコアでは，市内と県内市内の知名度の差の影響を大きくし，より局所的な知名度に重みをおいたローカルスコアについて，より近似できると示唆された．また，LS_0.25 の場合でも，nDCG の評価値は比較的高い数値であることから，ローカルスコアが高い実世界スポットについては，正しくスコアを推定できていると考えられる．

また，関西エリアと九州エリアでは，地名の地理的広さを用いない LS'_IDF と LS'_$GEOIDF$ が高い値となった．nDCG で比較すると，特に九州エリアでは，共起地名を用いない LS'_IDF が最も高い数値を示した．nDCG はランキングの上位に重みをかけた指標であるため，ローカルスコアが高いスポットについて適切に上位にランキングする点においては，知名度を利用しない手法のほうが精度が高い場合があるといえる．このことから，知名度の差を用いる方法に LS'_IDF や LS'_$GEOIDF$ を組み合わせることで，より精度の高い推定が可能になると考えられる．

さらに，どのエリアにおいても，LS'_T のように，県内市内と県外を市外として扱った場合でも，県内市外と県外を分けて知名度の差を用いた手法と比べ，P@15 および nDCG の値に差はみられない．一方，LS'_TC のように，県外の知名度を用いない場合には，知名度の差を用いた他の手法と比べると，P@15 と nDCG の値がともに低くなることが多い．このことから，市内，県内市外，県外のすべての知名度を用いることで，精度高く推定できることが示唆された．

LS'_RT と知名度の差を用いた手法に基づく推定ローカルスコアによるランキングを比較したところ，LS'_RT は LS'_0.75 と最も推定ローカルスコアが似ていることが分かった．LS'_RT と LS'_0.75 が等しいと仮定すると，

$$\begin{aligned} LS'\_RT &= LS'\_0.75 \\ &= 0.75(LS'\_RT - LS'\_RC) \\ &\quad + 0.25(LS'\_RC - LS'\_RT) \end{aligned}$$

となる．この式を整理すると，

$$LS'\_RT + 2LS'\_RC = LS'\_RP$$

となり，推定知名度 LS'_RT，LS'_RC，LS'_RP はすべて 0 以上の値となることから，

$$LS'\_RP \geq LS'\_RT$$
$$LS'\_RP \geq 2LS'\_RC$$

となる．このことから，LS'_RT と $LS'\_0.75$ の値が等しい場合には，各推定知名度には上記の関係が成り立つことが分かる．よって，今回用いた実世界スポットは，県外での推定知名度が高く，市内や県内市外と県外の推定知名度の差が大きくなるようなものが多かったと考えられる．これは，本実験でのデータ依存の問題であり，すべての実世界スポットにおいて，LS'_RT と知名度の差を用いた手法による推定ローカルスコアが同等になるとはいえない．よって，市内の知名度のみを用いることでローカルスコアが適切に推定できるとは限らない．一方，知名度の差を用いた手法であれば，どのような推定知名度の差を持つような実世界スポットでも汎用的に使えると考えられる．

次に，$LS\_0.75$ を正解とした場合についてエリア別に比較する．関東エリアでは，P@15 と nDCG のどちらの値においても LS'_IDF と LS'_GEOIDF と比較すると，知名度の差を用いた手法において正解率が高い数値を示した．その中でも，P@15 では，$0.5 \leq \omega < 1$ として知名度の差を用いた手法の値が，$\omega < 0.5$ として知名度の差を用いた手法に比べて高くなっているが，nDCG では，$\omega < 0.5$ として知名度の差を用いた手法の値が，$0.5 \leq \omega < 1$ として知名度の差を用いた手法より高くなった．ランキング上位を重視する nDCG において $\omega < 0.5$ として知名度の差を用いる手法の値が高くなったことから，ランキング上位の推定を重視する場合，ω を小さくすることで高精度に推定できると考えられる．関西エリアでは，P@15 の結果では，どの手法においてもあまり差がみられなかったが，nDCG の値で比較すると，知名度の差を用いた手法は，$LS'\_0.9$ で最も高い値となり，ω の値を大きくするにつれて値が高くなる傾向がみられた．九州エリアでは，P@15 の値では，$LS'\_0.25$ で最も高い値となった．nDCG の値では，知名度の差を用いた手法で比較的値が高く，なかでも ω の値が小さいほど高い値となった．このことから，関東エリアと同様に，ω の値を小さくするほど高精度に推定できたと考えられる．よって，関西エリアの京都市のように，県外からの観光客も多いエリアでは，市内と県内市外の知名度の差に重みをおき，関東エリアの横浜市のように県内外どちらからも観光客が多いエリアや，九州エリアの福岡市のように県外からの観光客が少ないエリアでは，県内市外と県外の知名度の差に重みをおくことで，より適切にローカルスコアを近似できることが示唆された．

また，ω の値が異なる 6 手法（$LS'\_0.1$, $LS'\_0.25$, LS'_TP, $LS'\_0.75$, $LS'\_0.9$, LS'_TC）について，P@15 と nDCG それぞれの全エリア平均値の推移を**図 5** に示す．ω の値が変わっても評価値にほとんど差がみられないことから，式 (2) によって求められるローカルスコアは，ω の値による影響は少なく，各区域での知名度の差に依存していることが分かった．次節でエリア別の詳細な分析を行う．

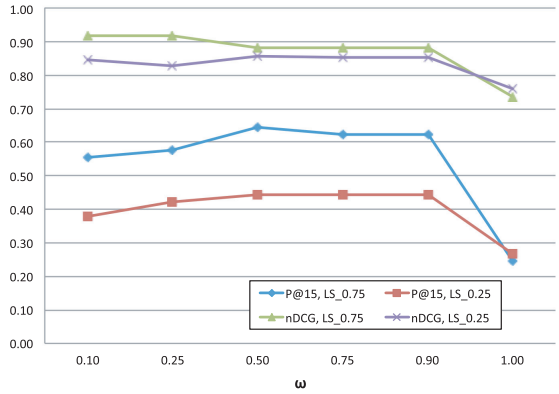

図 5 ω の値による P@15 と nDCG の平均値の推移

Fig. 5 The mean value of P@15 and nDCG at different ω values.

4.4 推定知名度の分析

提案手法で推定ローカルスコアのを求めるために用いた各区域での推定知名度が，評価データを用いて求められた知名度を正しく近似しているかどうか検証を行った．ここでは，各実世界スポットの区域別知名度を確率と見なし，評価データを用いた知名度を真の確率分布，推定知名度を比較対象の確率分布として Kullback-Leibler ダイバージェンス（以下，KLd）を用いることで，評価データによる知名度と推定知名度の差を検証した．なお，KLd は分布間の類似度として用いられ，値が低いほど比較対象の確率分布と真の確率分布の差が少なく，推定知名度が実際の知名度に近いことを表す．提案手法による知名度推定値と評価データによる知名度の KLd をエリア別に昇順に並べたものを**図 6** に示す．

関東エリア（図 6 (a)）では，KLd が高い実世界スポットとして「横浜メディアビジネスセンター」「横浜みなとみらいスポーツパーク」「赤い靴はいてた女の子像」など，正式名称が長いものが多く含まれていた．また「MotionBlue 横浜」のようにアルファベットと日本語が混ざったスポットも複数含まれていた．これらの実世界スポットについては，実世界スポットの正式名称を含む文書数が少ないため，本実験では LS'_IDF も高かった．これは，ブログ記事はニュース記事などとは異なり，実世界スポットの名称を正式に記述するよりも，投稿者がふだん呼び慣れている通称や略称などが使われることが多いことが原因と考えられる．そのため，提案手法の愚直な適用では，正式名称でブログ記事に書かれている数が少なければ，正しく知名度

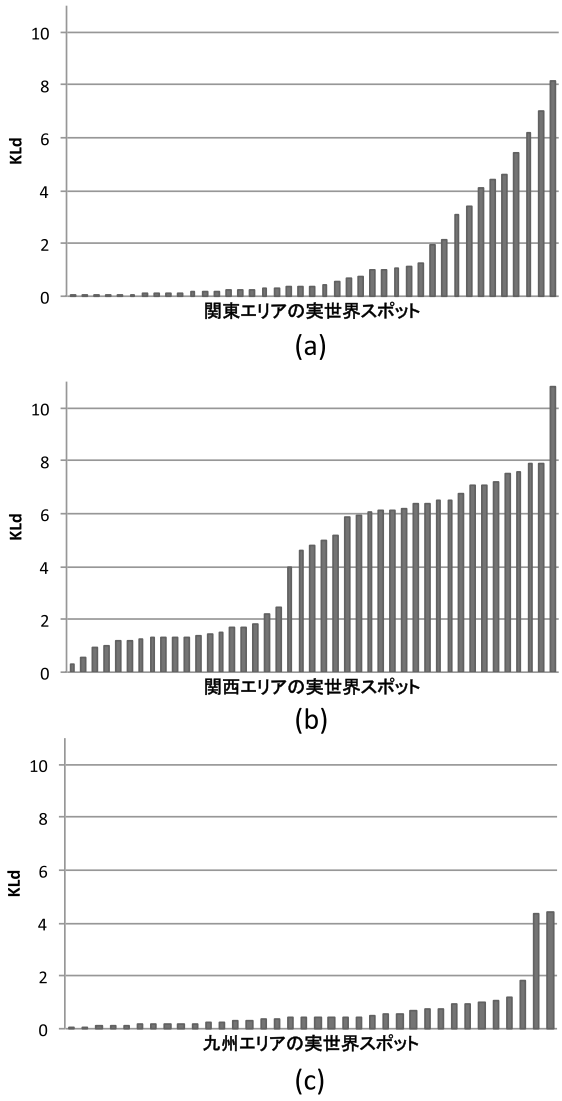

図 6 推定知名度の KLd
Fig. 6 KLd between estimates of name recognition.

が推定できない．これを解決するためには，ブログ記事を取得する際に通称や略称を考慮する必要がある．

次に，関西エリア（図 6 (b)）は他のエリアに比べると KLd が高い実世界スポットが多く見受けられる．実際に 44 個の実世界スポットのうち 26 個は KLd が相対的に高い値を示しており，そのほとんどは寺や神社であった．2.2 節で述べたとおり，今回拠点市として選んだ京都市は，府外からの観光客数が多い．そのため，府外の人でも京都の詳細な地名を知っている可能性も高く，実世界スポットに近い人のみが詳細な地名を使うという本研究の仮定から外れていると考えられる．このような場合，地名だけでなく，実世界スポットを言及する際に用いた固有名詞や言い回しなど，地元の人ならではの特徴語を用いることによって，より精度高く知名度を推定できると考えられる．

一方，九州エリア（図 6 (c)）では，他のエリアに比べると KLd が低い実世界スポットが多く，知名度を正確に近似できているといえる．2.2 節で述べたとおり，福岡市は，県外からの観光客の割合が比較的少ないことから，全国のテレビで取り上げられるような有名なスポットと，地元の人だけが知っているスポットが明確に分かれていると考えられる．このように，周辺と離れた地域で知名度が明らかに異なるようなエリアについては，提案した推定知名度によって実際の知名度を正確に推定することが可能であり，推定ローカルスコアも正確に近似できることが示唆された．

また，全エリアに共通して，実際の知名度の高さ以上に，ブログ記事に書かれやすいスポットがあると推測される．たとえば，関西の京都競馬場や関東のウインズ新横浜や横浜アリーナ，九州のマリンメッセ福岡やレベルファイブスタジアムなど，スポーツと関わりのある実世界スポットや，九州の HKT48 劇場や FBS 福岡放送などのメディアと関わりのある実世界スポットにおいては，スポット名を含む文書数も多く，LS'_IDF による推定ローカルスコアも実際より低くなった．このように，ウェブ上で話題になりやすい実世界スポットについては，実際のローカルスコアよりも低く推定されてしまうため，これを補正する手法を用いる必要がある．

これらの結果から，ローカルスコアをより高精度に推定するために，地名だけでなく他の特徴語なども考慮した推定ローカルスコアの計算手法の検討が必要であることが分かった．

5. 関連研究

この章では，本研究の関連研究について述べる．

5.1 実世界スポットの関連研究

実世界スポットに関する研究は大きく以下の 2 つに分けられる：

(1) 未知の実世界スポット抽出
(2) 既知の実世界スポットに対するアノテーション

(1) 未知の実世界スポット抽出とは，実世界スポットの位置情報と名称をウェブ上の情報源から抽出するタスクである．既存研究としては，Rattenbury ら [3] や Yang ら [4] が提案している，写真共有サイトに投稿された写真と写真にタグ付けされた位置情報などを用いて，イベントや場所を自動抽出する手法がある．また，Cheng ら [5] は，位置情報を持たないマイクロブログを用いて，記事中に書かれている内容から，どこから投稿された記事なのかを推定する手法を提案している．これを用いて推定した位置情報と記事に含まれる実世界スポット名を組み合わせることで，未知の実世界スポットについて名称と位置を抽出することも可能であると考えられる．

(2) 既知の実世界スポットに対するアノテーションとは，既知の実世界スポットについてウェブ上の情報源を用いて特徴を抽出し，情報を付加するタスクである．本研究は，既知の実世界スポットに対して，知名度の地理的広がりの

情報を付与するという観点から (2) の研究である．(2) の既存研究はいくつかある．Kurashima ら [6] は，ブログを用いて体験表現を判別し，抽出したランドマークについて，そこでできる体験に関する話題語を付与する技術を提案している．Fujisaka ら [7] や渡辺ら [8] は，実世界スポット周辺の位置情報が付与されたマイクロブログの投稿数を用いて，実世界スポットが人を集める人気スポットであるかどうかの判定をしている．Gao ら [9] は，Flickr [*7] の投稿や旅行サイトのユーザの口コミに基づき，観光客の間で人気の観光地を抽出し，ランキングする手法を提案している．これらは，実世界スポットがどれだけ人気かという情報を付与しているととらえることができる．

また，地域と実世界スポットの関係性を表す尺度を用いて情報を付与する技術もいくつかある．廣嶋ら [10] は，ある場所において特徴的なキーワードを獲得するため，共起する地名表現から語の分布を考慮した方法を提案している．Yin ら [11] は，位置情報付きの文書の緯度経度情報とテキストを用いて，ある場所で特徴的なトピックを見つけるモデルを提案している．ここで，キーワードやトピックを実世界スポットの名称とした場合，実世界スポットがある場所において特徴的かどうかという情報を付与しているととらえられる．Zhang ら [12] は，Flickr の写真に付与された位置情報と時間情報を特徴として，写真に付与されたタグをクラスタリングする手法を提案している．これらのタグの中には実世界スポットの名称も含まれており，実世界スポットについて，地理的時間的に特徴的かどうかという情報を付与しているととらえられる．奥ら [13] は，グルメ情報サイトの店名，緯度経度情報，PR 文を用いて，地域限定性という尺度に基づいて実世界スポットにアノテーションしている．地域限定性は，ユーザの地元では利用できず，旅行先や出張先などの現地でしか利用できないような実世界スポットほど値が高くなるような尺度である．また，手塚ら [14] は，あるキーワードについて，キーワードと地理空間への関連性を表す地域性という尺度に基づいてアノテーションしている．ここで，キーワードを実世界スポットの名称と置き換えると，実世界スポットの地理空間への関連性を地域性で表すことができると考えられる．Quercia ら [15] は，携帯電話の移動履歴とイベントのデータベースを組み合わせて，ユーザの居住地に応じたイベントを推薦する技術を提案しており，多くの人が訪れたイベント，居住地の近くで行われたイベントなど，実世界のイベントについて 6 つの尺度に基づくスコア付けをしている．ここで，イベントを実世界スポットと置き換えると，実世界スポットに訪れた人の居住地に基づく尺度でアノテーションされているととらえられる．

本手法では，実世界スポットについて，知名度の地理的広がりを表す地域局所性という従来手法には用いられなかった新しい軸によるアノテーション手法を提案している．また，ブログ記事を用いて地域局所性を自動推定する手法であり，ここで推定した地域局所性は他の軸と組合せ可能であると考えられる．既存手法によってアノテーションされた実世界スポットに対し，さらに地域局所性に応じた情報を付与することで，実世界スポットが存在する周辺のみで知られている実世界スポットと，全国的に有名な実世界スポットというように，知名度の地理的広がりによる判別が可能となる．

5.2 投稿者の居住地推定

Yamaguchi ら [16] は，居住地が近い友人と SNS 上で多く結びついているユーザをランドマークと定義し，ユーザの居住地やプロフィールを推測している．Sadliek ら [17]，Li ら [18]，McGee ら [19] は，Twitter の友人関係と投稿内容を用いて，ユーザの家の位置を予測する手法を提案している．また，Kinsella ら [20] は，位置情報付きのマイクロブログを用いて，エリアごとの言語モデルを生成する手法を提案しており，これを用いて郵便番号レベルでユーザの居住地を予測している．Chandra ら [21] は，SNS 上でのユーザの会話を基に，ユーザの位置を都市レベルで推定する手法を提案している．これらの手法を用いて，実世界スポットについて言及しているブログ記事やツイート記事の投稿者の居住地が分かれば，より高い精度で各区域での知名度を推定することができると考えられる．我々は，投稿者と実世界スポットの関係は，現在の居住地だけでなく，過去の居住地や学校・職場があるエリアなども影響すると考えている．本研究で用いた，言及する際に用いる地名の地理的広さは，このような投稿者と実世界スポットの関係も適切に表すことが可能である．

6. まとめ

本研究では，住所を持ち，ユーザの訪問対象である実世界スポットに着目し，実世界スポットの知名度広がりに基づく地域局所性を定量化したローカルスコアの計算方法を定義した．居住地ごとの被験者評価データを用いたローカルスコアにより，実世界スポットごとに知名度の広がり方の傾向が異なることを確認した．また，ブログ記事においてスポット名と文書内共起する地名の地理的広さを用いて，各実世界スポットの区域別知名度を推定し，ローカルスコアの推定手法を提案した．評価実験を通じて，スポット名だけでなく共起する地名を用いたり，その地名の地理的広さを考慮したりするなどによって，高精度にローカルスコアを推定することができることが分かった．これにより，被験者評価データを利用しなくても一定の精度で任意の実世界スポットの地域局所性推定が可能となり，たとえば地域情報サービスのパーソナライズなどに活用できると考え

*7 https://www.flickr.com

られる．今後の課題としては，実世界スポットがあるエリアの特徴や実世界スポットそのものの特徴を考慮したアプローチに取り組むことがあげられる．今回は実世界スポットが存在する住所に現れる地名を県・市・町の 3 レベル別に分けて扱ったが，同レベルの地名が表す地域でも面積や人口などの特徴が異なる．各地名のレベルだけでなく，その地域の特徴も考慮することによって，知名度の広がり方をより詳細に表現できる可能性がある．また，地名以外の地域特徴語を組み合わせた推定手法の検討が必要である．

参考文献

[1] 平野　徹，松尾義博，菊井玄一郎：地理的距離と有名度を用いた地名の曖昧性解消，全国大会講演論文集，Vol.70, No.2, pp.2:85–2:86 (2008).

[2] Kalervo, J. and Jaana, K.: Cumulated gain-based evaluation of IR techniques, *ACM Trans. Inf. Syst.*, Vol.20, No.4, pp.422–446 (2002).

[3] Rattenbury, T., Good, N. and Naaman, M.: Towards automatic extraction of event and place semantics from Flickr tags, *Proc. SIGIR '07*, pp.103–110 (2007).

[4] Yang, Y., Gong, Z. and U, L.H.: Identifying points of interest by self-tuning clustering, *Proc. SIGIR '11*, pp.883–892 (2011).

[5] Cheng, Z., Caverlee, J. and Lee, K.: You are where you tweet: a content-based approach to geo-locating Twitter users, *Proc. CIKM '10*, pp.759–768 (2010).

[6] Kurashima, T., Tezuka, T. and Tanaka, K.: Mining and visualization of visitor experiences from urban blogs, *Proc. DEXA '07*, pp.213–222 (2007).

[7] Fujisaka, T., Lee, R. and Sumiya, K.: Discovery of user behavior patterns from geo-tagged micro-blogs, *Proc. ICUIMC '10*, pp.36:1–36:10 (2010).

[8] 渡辺一史，大知正直，岡部　誠，尾内理紀夫：Twitter を用いた実世界ローカルイベントの検出，第 4 回楽天研究開発シンポジウム (2011).

[9] Gao, Y., Tang, J., Hong, R., Dai, Q., Chua, T.-S. and Jain, R.: W2Go: A travel guidance system by automatic landmark ranking, *Proc. MM '10*, pp.123–132 (2010).

[10] 廣嶋伸章，安田宜仁，藤田尚樹，片岡良治：地理情報検索におけるクエリ入力支援のための特徴語の提示，人工知能学会全国大会 (2012).

[11] Yin, Z., Cao, L., Han, J., Zhai, C. and Huang, T.: Geographical topic discovery and comparison, *Proc. WWW '11*, pp.247–256 (2011).

[12] Zhang, H., Korayem, M., You, E. and Crandall, D.J.: Beyond co-occurrence: discovering and visualizing tag relationships from geo-spatial and temporal similarities, *Proc. WSDM '12*, pp.33–42 (2012).

[13] 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，情報処理学会論文誌　データベース，Vol.5, No.3, pp.97–116 (2012).

[14] 手塚太郎，近藤浩之，田中克己：混合ガウス分布を用いたウェブコンテンツの地域性推定とオブジェクトレベルローカルサーチ，情報処理学会論文誌 データベース，Vol.1, No.1, pp.13–25 (2008).

[15] Quercia, D., Lathia, N., Calabrese, F., Lorenzo, G.D. and Crowcroft, J.: Recommending social events from mobile phone location data, *Proc. ICDM '10*, pp.971–976 (2010).

[16] Yamaguchi, Y., Amagasa, T. and Kitagawa, H.: Landmark-based user location inference in social media, *Proc. COSN '13*, pp.223–234 (2013).

[17] Sadilek, A., Kautz, H. and Bigham, J.P.: Finding your friends and following them to where you are, *Proc. WSDM '12*, pp.723–732 (2012).

[18] Li, R., Wang, S., Deng, H., Wang, R. and Chang, K.C.-C.: Towards social user profiling: Unified and discriminative influence model for inferring home locations, *Proc. KDD '12*, pp.1023–1031 (2012).

[19] McGee, J., Caverlee, J. and Cheng, Z.: Location prediction in social media based on tie strength, *Proc. CIKM '13*, pp.459–468 (2013).

[20] Kinsella, S., Murdock, V. and O'Hare, N.: I'm eating a sandwich in Glasgow: Modeling locations with tweets, *Proc. SMUC '11*, pp.61–68 (2011).

[21] Chandra, S., Khan, L. and Muhaya, F.B.: Estimating Twitter user location using social interactions–A content based approach, *Proc. SocialCom/PASSAT '11*, pp.838–843 (2011).

推薦文

本稿では，地域局所性の推定という，重要かつ面白い課題に取り組んでいます．地域局所性を評価するために，比較的大規模なユーザ調査を行い各地域における実世界スポットの知名度を評価，分析している点が高く評価できます．また，各地域における知名度を推定するための，地名の言及の仕方に着目した手法も，仮説の検証と結果の議論が丁寧になされており，完成度の高い論文であると思います．以上の理由により，当該論文を，論文誌推薦に値する論文であると判断いたしました．

（FIT2013 第 12 回情報科学技術フォーラム
プログラム委員長　荒川賢一）

徳永 陽子　（正会員）

2008 年同志社大学工学部知識工学科卒業．2010 年京都大学大学院情報学研究科社会情報学専攻修士課程修了．同年日本電信電話株式会社入社．現在，NTT サービスエボリューション研究所に所属．主として情報検索に関する研究開発に従事．

数原 良彦 （正会員）

2006 年慶應義塾大学理工学部管理工学科卒業．2008 年同大学院理工学研究科開放環境科学専攻修士課程修了．同年日本電信電話株式会社入社．現在，NTT サービスエボリューション研究所に所属．主として情報検索，機械学習に関する研究開発に従事．人工知能学会，言語処理学会各会員．

佐藤 吉秀 （正会員）

2000 年京都大学工学部電気電子工学科卒業．2002 年同大学院情報学研究科システム科学専攻修士課程修了．同年日本電信電話株式会社入社．現在，NTT サービスエボリューション研究所に所属．以来，Web マイニングの研究開発に従事．

戸田 浩之 （正会員）

1997 年名古屋大学工学部材料プロセス工学科卒業．1999 年同大学院工学研究科材料プロセス工学専攻博士課程前期課程修了．同年日本電信電話株式会社入社．現在，NTT サービスエボリューション研究所に所属．以来，情報検索，データマイニングの研究開発に従事．2007 年筑波大学大学院システム情報工学研究科コンピュータサイエンス専攻博士後期課程修了．博士（工学）．ACM，日本データベース学会各会員．

鷲崎 誠司 （正会員）

1988 年名古屋大学理学部数学科卒業．同年日本電信電話株式会社入社．自然言語処理，情報検索に関する研究開発に従事．現在，NTT メディアインテリジェンス研究所主幹研究員．