RD-003

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
Locality inference of real-world spots considering the extent of name recognition

田中 陽子 † 数原 良彦 † 佐藤 吉秀 † 戸田 浩之 † 鷲崎 誠司 †
Yoko Tanaka Yoshihiko Suhara Yoshihide Sato Hiroyuki Toda Seiji Susaki

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挙げられる．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．例えば，全国的によく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出来れば，ユーザが新たな目的地を発見することができる．例えば，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既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．

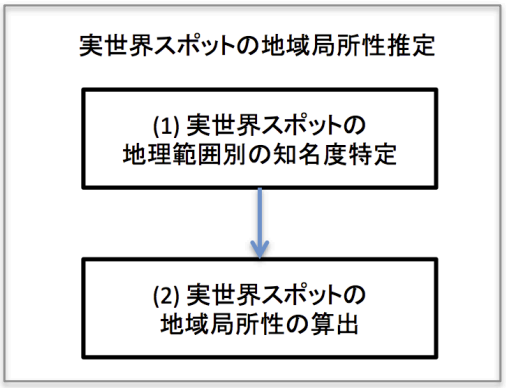

図 1: 本稿の流れ

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．

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，様々な産業が盛んな九州エリアの 3 つを拠

†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

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

Copyright © 2013 by Information Processing Society of Japan and
The Institute of Electronics, Information and Communication Engineers
All rights reserved.

表 1: 評価に用いたエリア

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: 被験者数（人）

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

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」，拠点府県に隣接する都道府県を「隣接県」，隣接県を除く他の府県を「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，京都市から 57 個，福岡市から 40 個，横浜市から 53 個，計 150 個を選定した．

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，例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等も考慮した．エリア別・区域別の被験者数は表 2 に示す．

2.2. 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わかる．

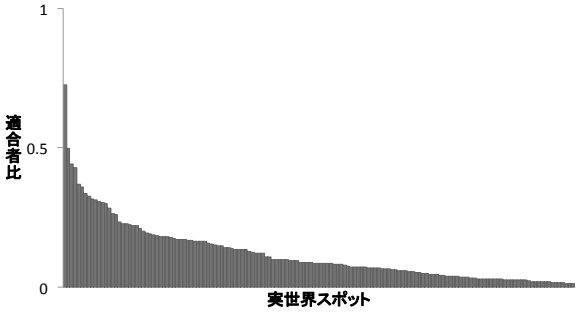

図 2: 各実世界スポットの適合者比

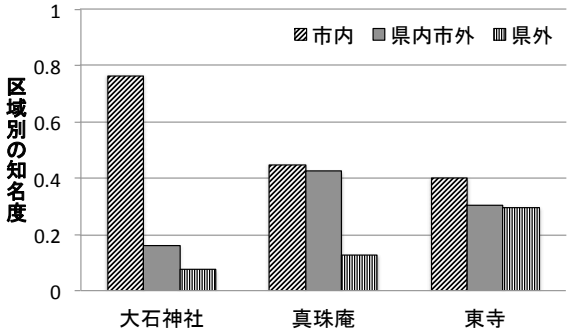

図 3: 各スポットの区域別知名度

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 に挙げた 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

Copyright © 2013 by Information Processing Society of Japan and
The Institute of Electronics, Information and Communication Engineers
All rights reserved.

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，上記で例に挙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，県内市外と県内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わかる．

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出来る．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．次の章では，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．

2.3. ローカルスコアの定義

評価データの分析結果を踏まえて，市内と県内市外の知名度の差を県内市外と県外の知名度の差よりも重視するように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．

$$LS\,(s) = \lambda(RT_s - RC_s) + (1-\lambda)\,(RC_s - RP_s) \tag{1}$$

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．また，前述のとおり市内と県内市外の差が地域局所性への影響が大きいことから，λ>0.5 であることが望ましい．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．本稿では $\lambda = 0.75$ とする．

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 3 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出来るようになった．

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[3)] では，筆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．プロファイルに筆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．

すべての記事に対して，筆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，その記事の筆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．筆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，筆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，筆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

[3)] http://twitter.com

Copyright © 2013 by Information Processing Society of Japan and
The Institute of Electronics, Information and Communication Engineers
All rights reserved.

表 3: ローカルスコアによるランキング

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

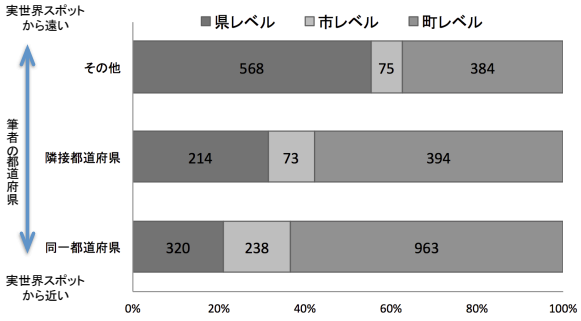

図 4: 筆者の居住区域と共起地名の関係

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わかった．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一つであると考えられる．例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，(1) 式で用いた知名度 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，常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，筆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，

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

Copyright © 2013 by Information Processing Society of Japan and
The Institute of Electronics, Information and Communication Engineers
All rights reserved.

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．

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もと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，独自に収集した約 8,000 万件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．

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 [5] を用いた．このとき，地名は後方一致のみを見ることとし，例えば「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，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 2 手法と，推定知名度を用いる 4 手法の計 6 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．

$LS'\_0.75$: (1) 式で，$\lambda = 0.75$ としたもの．

$LS'\_0.9$: (1) 式で，$\lambda = 0.9$ としたもの．

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

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

推定知名度を利用しない 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，筆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，地名共起の有無に関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，推定知名度を用いた 4 手法について述べる．$LS'\_0.75$ と $LS'\_0.9$ は (1) 式の λ の値を変えたもので，$LS'\_0.9$ の方が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，推定に用いる地名のレベルを絞った手法である．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．

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 文書とみなすと，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 点とした時の nDCG[2].

の 2 つを用いた．

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 4 に示す．まず，P@15 の値について述べる．関東エリアでは，LS'_IDF と LS'_GEOIDF と比較すると，$LS'\_0.75, LS'\_0.9, LS'\_TC$ といった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．関西エリアでは，知名度の差を利用しない 2 手法は 3 割程度の正解率に留まった一方，知名度の差を用いた 4 手法はいずれも 5 割以上の正解率となった．九州エリアでも，知名度の差を用いた 4 手法ではいずれも 7 割を超える高い数値となった．いずれのエリアでも，P@15 の値が知名度の差を用いた手法で高かったことから，知名度の差を用いることによって正解の実世界スポットをより多く当てることができると示された．

次に，nDCG の値について述べる．関西エリアでは，P@15 と同様に知名度の差を用いた 4 手法の数値が高かった．一方，関東エリアでは，地名の地理的広さを用いない LS'_GEOIDF が最も高い数値を示した．また，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

Copyright © 2013 by Information Processing Society of Japan and
The Institute of Electronics, Information and Communication Engineers
All rights reserved.

表 4: ランキング比較評価

	P@15			nDCG		
	関東	関西	九州	関東	関西	九州
LS'_IDF	0.267	0.333	0.467	0.780	0.891	**0.975**
LS'_GEOIDF	0.267	0.333	0.467	**0.835**	0.875	0.919
$LS'\_0.75$	**0.467**	**0.533**	**0.733**	0.804	**0.905**	0.941
$LS'\_0.9$	**0.467**	**0.533**	**0.733**	0.804	**0.907**	0.937
LS'_TC	**0.467**	**0.533**	**0.733**	0.805	**0.908**	0.935
LS'_TP	**0.467**	**0.533**	**0.800**	0.800	0.897	**0.946**

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．

このように，エリアによって評価結果が異なるものの，どの手法においても P@15 の値は知名度の差を用いた 4 手法で高かったことから，知名度の広がりを考慮することでローカルスコアの推定精度に寄与できたといえる．一方，nDCG の結果を受けて，スポット名の出現文書数や地理的広さを考慮しない地名共起文書数による推定を組み合わせることで，より推定精度を向上できる可能性が示唆された．次章にてエリア別の詳細な分析を行う．

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みなし，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 5 に示す．

関東エリア (図 5(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，筆者が普段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 5(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．今回拠点市として選んだ京都市は国内有数の観光名所であり，府外からも多くの観光客が訪れる．また，リピーターも多く，京都の観光だけに特化したブログサイトも多い．これらのことから，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．また，関東エリアの拠点市である横浜市でも，同様の性質があると考えられる．これは，横浜市が首都圏である上，近隣府県との交通網も発達しており，県外に住んでいても市内と行き来を繰り返す人が多いからである．

一方，九州エリア (図 5(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言える．福岡市も九州の中心都市ではあるものの，他の 2 つのエリアと比較すると，地元の人だけが知っているスポットと全国的に知られ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．例えば，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．

Copyright © 2013 by Information Processing Society of Japan and
The Institute of Electronics, Information and Communication Engineers
All rights reserved.

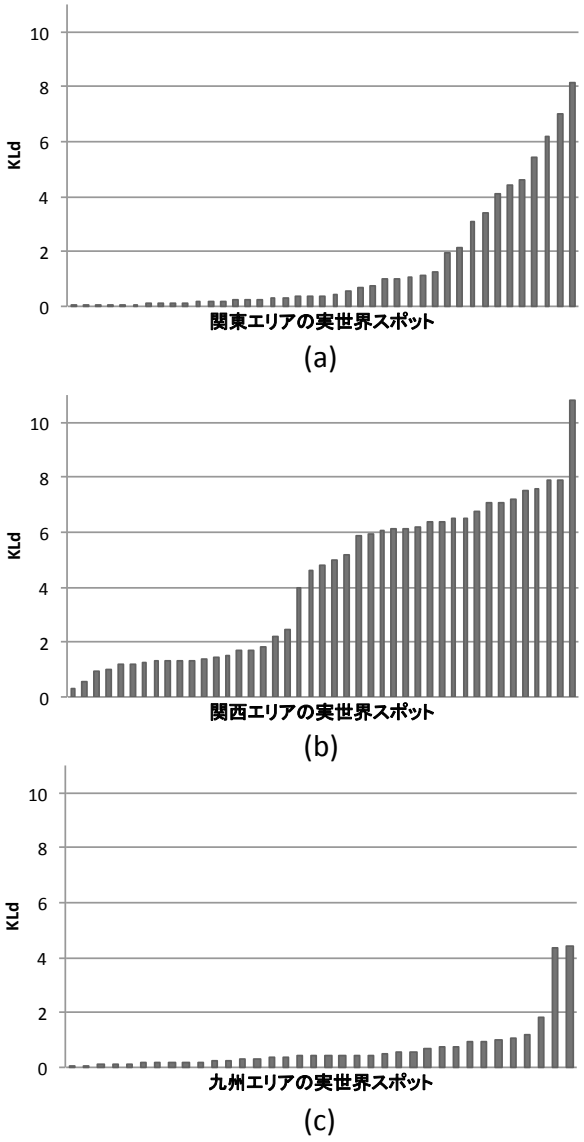

図 5: 推定知名度の KLd

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わかった．

5. 関連研究

実世界スポットに関する研究は，これまでにも多くの研究が行われている．Fujisaka ら [1] は，位置情報付きのマイクロブログを用いて，ある実世界スポットの周辺の位置情報を含む記事が多く投稿された場合，そこは人が集まる人気の実世界スポットであると判定することを試みている．また，渡辺ら [4] も同様に，ある期間に多くのユーザから位置情報付きで記事が投稿された場所を人気スポットとして，関連する情報とともに抽出する技術を提案している．これらの技術は，集まった人の居住区域を区別しておらず，地域局所性の推定には適用できないと考えられる．また，前述のとおり，位置情報では地名の地理的広さを考慮することができないため，本研究には適用できないと考えられる．

廣嶋ら [6] は，共起する地名表現から語の分布を考慮し，場所に関する特徴的なキーワードを獲得する方法を提案している．また，倉島ら [3] は，ブログを用いて体験表現を判別し，ランドマークと話題語を抽出する技術を提案している．どちらもブログ記事内の地名を用いている点では本手法と共通しているが，キーワードやランドマークの知名度による区別は行なっていない．これらの手法と本手法を用いて，共起している地名の広さを用いて知名度の広がりを推測することで，ユーザのニーズに合わせて異なるキーワードや話題語の提示が可能になると考えられる．

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．居住地毎の被験者評価データを用いたローカルスコアにより，実世界スポット毎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わかった．これにより，被験者評価データを利用せずと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，例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挙げられる．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] Tatsuya Fujisaka, Ryong Lee, and Kazutoshi Sumiya. Discovery of user behavior patterns from geo-tagged micro-blogs. pp. 36:1–36:10, 2010.

[2] Järvelin Kalervo and Kekäläinen Jaana. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, Vol. 20, No. 4, pp. 422–446, October 2002.

[3] 倉島健, 手塚太郎, 田中克己. Blog からの街の話題抽出手法の提案. 電子情報通信学会第 16 回データ工学ワークショップ, 2005.

[4] 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.

Copyright © 2013 by Information Processing Society of Japan and
The Institute of Electronics, Information and Communication Engineers
All rights reserved.

[5] 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, No. 2, pp. 2–85, 2008.

[6] 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.

Copyright © 2013 by Information Processing Society of Japan and
The Institute of Electronics, Information and Communication Engineers
All rights reserved.