情報処理学会論文誌　データベース　Vol. 1　No. 1　13–25 (June 2008)

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

手塚太郎[†1]　近藤浩之[†2]　田中克己[†2]

近年，ウェブ上で多数の地域情報検索エンジン（ローカル検索エンジン）が実用化されているが，それらは主にウェブページを検索単位としている．一方，任意の概念に対し，関連するウェブページの内容を集約してユーザに提示する「オブジェクトレベルサーチ」がさかんに研究されている．本研究では「オブジェクトレベルローカルサーチ」の実装として，任意のキーワード型検索クエリに対し，実時間でウェブ検索結果を解析し，クエリが関連する地域を求めるシステムについて述べる．関連地域を取得するにあたって，検索結果中に現れる地物の座標から混合ガウス分布を推定し，規模の小さなクラスタをフィルタリングすることで，適切なマッピングが得られることを示す．推定された関連地域は地図インタフェース上で視覚化される．

Estimation of Relevant Regions for Web Content by Gaussian Mixture Models for Object Level Local Search

Taro Tezuka,[†1] Hiroyuki Kondo[†2] and Katsumi Tanaka[†2]

The rapidly increasing use of geographically restrictive web search engines has made determining the geographic relevance of web content an important task. Most of these search engines are aimed at finding web pages that are relevant to specific geographic area. On the other hand, there is a growing demand for "object level search," which enables users to obtain integrated information on an arbitrary topic, rather than looking through individual pages. In this paper, we present a method for "object level geographic search" that estimates geographic regions that an arbitrary topic is relevant to. Our method uses Gaussian mixture models to estimate the geographic relevance of web pages and arbitrary objects. We have also implemented a visualization interface based on our proposed method. The system uses a set of web services and estimates relevance in real time.

1. はじめに

特定の空間的領域を対象としてウェブ検索を行う地域情報検索（ローカル検索）のシステムが近年，大きな注目を集めている．多数の大手検索エンジンがローカル検索サービスを提供し，さらに，それらの機能を提供する API が一般に公開されていることによって，数多くの応用システムが開発されてきた[22)–24)]．しかし，これらのサービスによって検索されるのはウェブページであり，ユーザは個別のページの閲覧を通してしか地域的な情報を得ることができない．

一方，ウェブ検索の分野において近年，多数のページに記述されている内容を解析することで，任意の概念に関する知識を取得する「オブジェクトレベルサーチ」の研究がさかんになってきている．本研究で扱われる「オブジェクト」とは，ユーザがそれに関する情報を得たいと考える対象を指し，一般名詞と固有名詞をともに含むものである．研究によっては「テーマ」や「トピック」といった用語が使用されることもある．これに対し，「ページ」とはブラウザで閲覧される際に 1 つのまとまりとして扱われるウェブ情報の単位を意味する．ページが情報の提供形態に根ざした単位であるのに対し，オブジェクトはユーザの要求内容に基づく情報の単位である．オブジェクトレベルサーチのシステムにおいては，ユーザは多数のページを閲覧することなく，関心のある対象に関する知識を迅速に得ることができる．

本研究では地域情報を対象としたオブジェクトレベルサーチを実現させるため，ウェブページやオブジェクトが持つ空間的属性である「地域性」を実時間で推定する手法を述べる．地域性はウェブにおける地域情報検索を特徴付ける重要な性質であり，推定によって得られた地域性は様々なアプリケーションによって利用可能である．

本手法によって実現されるシステムの例として，ユーザが旅行先の都市において「紅葉」が有名な場所を調べたいという場合，地図インタフェース上で範囲を指定し，「紅葉」というオブジェクト名を入力することで，その範囲内で最も「紅葉」と関連の深い地域を表示させるというものがあげられる．

実時間で応答するシステムを目標としているため，検索エンジンが提供するウェブページ

†1 立命館大学情報理工学部
College of Information Science and Engineering, Ritsumeikan University
†2 京都大学大学院情報学研究科
Graduate School of Informatics, Kyoto University

の要約（スニペット）や，上位にランクされた少数のページから情報を抽出することが望まれる．そのため，ウェブページの位置推定に広く用いられている住所表記の抽出ではなく，文書中に現れる一般の地物名を推定用の基礎データとして使用する．このデータを用いる場合，住所表記を使用する場合と比較して，地物名が持つ曖昧性への対応が課題となる．本研究では地理空間上に広がる確率密度関数を導入することで，この問題に取り組む．

また，本研究で提案された手法をもとに，任意のオブジェクトに関連する空間的領域を地図インタフェース上で視覚化するシステムの実装を行った．

本論文の構成は以下のとおりである．2 章では関連研究について述べる．3 章では地域性のモデルを説明する．4 章では地域性の推定手法について述べる．5 章では実装について述べる．6 章では評価実験の結果を示す．7 章でまとめを行う．

2. 関 連 研 究

本章ではオブジェクトレベルローカルサーチならびに地域性を使用した検索に対する関連研究を述べる．

2.1 ローカルウェブサーチ

ウェブページのローカル検索に関しては，現在までに様々な研究が行われている．例として，McCurley はページ中に記述された住所・郵便番号・IP アドレス等に着目し，ウェブページを地理空間に位置づけるシステムを開発した[1]．Gao らはクローラの設計にあたり，URL やアンカータグ，ページ本体に含まれる地名や IP アドレスを利用し，特定の地域に偏らずにページを収集する手法を開発した[2]．Matsumoto らはウェブページの地域性が様々な側面を持つことを指摘し，それらの抽出手法について論じた[4]．さらに，ローカルウェブ検索においてどのようなインデックス構造が望ましいかに関して，R*木等を用いる提案も行われている[3]．これらの研究はウェブページを対象としており，オブジェクトレベルでの属性抽出や検索は扱われていない．また，主にクローリングによって得られたページへの処理を念頭に置いており，短い要約の集合から適切に範囲を推定するという課題は扱っていない．

2.2 オブジェクトレベルローカルサーチ

多数のページの集約によって得られた情報を提示する「オブジェクトレベルサーチ」は次世代の情報検索として注目されており，活発に研究が行われている．Nie らの研究ではページ集約によって得られる情報をウェブオブジェクト（web objects）と呼び，文書解析を用いて属性情報を自動抽出する研究を行っている[6),7]．オブジェクトレベルローカルサーチに関連しても，いくつかの先行研究がある．Buyukkoten らは新聞のウェブサイトにリンクしているページが空間的にどのように分布しているかを視覚化する研究を行った[8]．彼らの手法ではリンク元ページの位置を推定するのに IP アドレスを使用しているが，本研究のように日本国内を対象にする場合，IP アドレスの分布は大都市に集中し，また，ページ作成者やページの記述内容との関連も薄いと考えられるため，そのままでは使えないと考えられる．森本らは任意の単語をクエリとしてクローリングを行い，住所表記から座標を取得するシステム GMPSearch を開発した[9]．地域性の推定にはページ中に含まれる住所表記を利用している．そのため，住所が現れるだけの分量の文書数が必要であり，任意のクエリに対して実時間で応答するシステムにはなっていない．本研究では地物名一般を使用したことにより，少数の文書からも情報を引き出すことができ，ウェブ検索エンジンとの組合せによって実時間で応答するシステムが可能となっている．

2.3 ジオコーディング

地理情報システム（GIS）のデータを利用し，文書中に現れる住所あるいは地物名から地理空間上の座標を求める操作を**ジオコーディング**と呼ぶ．特に住所の表記揺れの問題に関して多くの研究が行われている．相良らはウェブページ中に含まれる住所表記を対象とし，ジオコーディングシステムの研究と実装を進めた[10]．また，住所以外では河川名等を対象として，文字列マッチングの手法を用いて地名データベースと照合させる研究を行っている[11]．有川は住所表記やタウンページのデータを用いたウェブコンテンツのジオコーディングに関してまとめている[12]．しかし，これらの研究は個々の地物名を座標にマッピングすることを目的としており，文書中に現れる地物名集合の空間的分布を推定するという考察は行われていない．

2.4 GIS 検 索

地理情報システムの応用として，空間的クエリの改善に関する研究が行われている．石川は位置情報が曖昧であるオブジェクトに対し，その位置を確率密度関数でモデル化し，空間データベース上で問い合わせる手法を提案した[13]．この研究で使用されている確率密度関数は多次元ガウス分布であり，本研究で用いられている混合ガウス分布の特殊なケースであるといえる．さらに，平均値ベクトル μ や分散共分散行列 Σ はあらかじめ与えられているとされており，本研究のようにデータからの学習は行われていない．また，物理空間上に存在し，位置に関して曖昧性を持つオブジェクトへの問合せを目的としたものであり，本研究で対象とするような一般的なオブジェクトの関連範囲を推定することを目的としたものではない．曖昧な境界線を持つ地理オブジェクトを表現するため，メンバシップ関数を利用す

る研究も行われている[14),15)]．しかし，これらは主に曖昧性の定式化を目的とした研究であり，本研究のような具体的な応用をともなったものではない．

3. 地域性のモデル

本研究ではウェブページならびにオブジェクトに対する地域性のモデルを提案する．

3.1 確率密度関数の利用

一般に，地域性は地理空間への関連を表す関数を意味する．特定の地域に強く関連するという性質や，特定の国と関連するという性質，あるいは特定の地点に関連するという性質は，いずれも地域性の一種である．

地域性の最も単純なモデルとしては，ある対象（ウェブページやオブジェクトを含む）に関して，地理空間上の各点ごとに関連度が決まるという構成が考えられる．関連度（relevance）に関する議論は文献 17) に準拠するが，任意の実数値を表すこととする．このような性質を持つ道具立ての 1 つは，ポテンシャル関数である．数学的扱いを容易にさせるため，値をすべて非負とし，空間全体で積分した際に 1 になるように正規化したものは確率密度関数である．そこで本研究では地域性を**地理空間上の確率密度関数**としてモデル化する．

確率モデルの 1 つの利点として，地物名の曖昧性によって真の地域性と無関係な座標データが現れることを確率的事象ととらえられる点があげられる．そのため，確率モデルに基づくフィルタリングによってノイズを取り除くことが可能になる．

3.2 混合ガウス分布

本節では地域性がどのような性質を持つかに関する議論をもとに，望ましい確率モデルに関して検討を行う．1 つめの性質として，地域性には複数の中心的領域が存在しうることがあげられる．たとえば「銘菓ひよ子」は福岡と東京，両方の土産物として知られているため，これらの 2 領域に強く関連する．また，それぞれの関連領域には中心（ピーク）があり，周辺に向かって関連度は減衰していくと考えられる．さらに，減衰は方向性を持つことがあり，等方的とは限らない．これらの要件は以下のようにまとめられる．

地域性の要件

- 空間上の座標を変数，関連度の大きさを値として持つ関数である．
- 関連度の大きさに関して，複数のピークを持てる．
- ピークから離れるにつれて，関連度が減衰する．
- 関連度の減衰は等方的とは限らない．

以上の条件を満たすものとして，混合ガウス分布（式 (1)）がある．混合ガウス分布は多次元ガウス分布の重ね合わせであり，空間全体で積分した際に 1 になるよう正規化したものである．

$$P(x) = \sum_{i=1}^{n} \alpha_i N(x; \mu_i, \Sigma_i), \qquad \sum_{i=1}^{n} \alpha_i = 1 \tag{1}$$

混合ガウス分布を構成する個々の d 次元ガウス分布 $N(x; \mu_i, \Sigma_i)$ は式 (2) によって表される．

$$N(x; \mu_i, \Sigma_i) = \frac{\exp(-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i))}{(2\pi)^{\frac{d}{2}} |\Sigma_i|^{\frac{1}{2}}} \tag{2}$$

ただし，$|\Sigma|$ は行列 Σ の行列式を表す．

ガウス分布は最も基本的な確率分布の 1 つであり，扱いの容易さ，統計的に好ましい性質等から幅広く使用されている．そこで本研究では地理空間を 2 次元と見なし，2 次元混合ガウス分布を確率モデルとして採用する．

混合ガウス分布はクラスタリング手法の一種ととらえることもできる．実際，本提案手法では混合ガウス分布の推定を行った後，標本（座標データ）を個々のガウス分布に振り分けるため，クラスタリングが行われているといえる．

他のクラスタリング手法と比較した場合，混合ガウス分布は本研究で扱われる課題においていくつかの優位性を持っている．まず，代表的なクラスタリング手法である K-means 法あるいは階層的クラスタリングを行った場合，互いに近接する要素どうしをまとめてクラスタが生成される．このとき，個々の要素がどの程度の確かさでクラスタに属しているかといった情報はクラスタリング結果には残らない．一方，混合ガウス分布の場合，クラスタが重なり合った状態も表現でき，それぞれの属性値においてクラスタへの関連度が定量的に表現されている．そのため，たとえばオブジェクトが県全体とも関連しているが，その中でも特にある市との関連が強い，といった状況を表せる等，表現力に富む．実際，混合ガウス分布の推定は K-means 法を一般化したものととらえることができる．

さらに，K-means 法や階層的クラスタリングにおいては個々のクラスタの重要性の度合いが明確な形では与えられないが，混合ガウス分布ではクラスタ c_i の重要性が重ね合わせの際の係数 α_i として，確率的に意味を持った値として得られる．そのため，要素数の少ないクラスタをフィルタリングする際等に確率的な基準を設けることができるといったメリットがある．

なお，混合ガウス分布の持つ制約の 1 つとして，個々のガウス分布が持つ等確率密度線

の形状が楕円に限られるという点があげられる．これは分布の表現力の限界を意味するが，本研究の実装では地図の縮尺の変更というインタラクティブな操作によって段階的に詳細な関連地域を求めていけるようにしているため，初期の推定においては関連地域の正確な形状は必ずしも重要ではないと考え，むしろ統計的に扱いやすいガウス分布を使用することが好ましいと判断した．

4. 地域性の推定手法

本章ではウェブページ/オブジェクトに対する地域性推定の手法について述べる．全体の流れは**図 1** に示した．

4.1 ページ集合からの地物名の取得

ウェブページを対象とする場合，形態素解析を適用し，地物名と分類された単語を収集する．オブジェクトを対象とする場合は，オブジェクトの名称をクエリとしてウェブ検索を行い，得られた検索結果を利用する．数多くの検索エンジンが検索用 API を提供しているため，これらの検索サービスをシステムに組み込んで実行することができる．

さらに，多くのウェブ検索 API では検索結果の要約文字列が提供されている．これらの文字列はスニペットとも呼ばれ，検索クエリに含まれる単語とマッチする部分の周辺を切り出した 100 文字前後のテキスト情報である．これらを利用する場合，実際の検索結果ページへのアクセスが不要になり，ページ本体よりも高速で取得できるため，本論文で述べる実装では積極的にこの情報を利用する．

4.2 ジオコーディング

ジオコーディングのステップでは，地物名を GIS を用いて座標に変換する．通常，ウェブページを座標に対応付ける研究においては，住所表記が使用されることが多い．詳細な住所が与えられている場合，その曖昧性は低く，空間へのマッピング精度は高い．しかし，特定の地域に関連性を持つウェブページのうち，住所表記を含むものは一部を占めるにすぎない．検索エンジンの上位結果を利用したシステムを構築する場合，事前に大量のページを集めてバッチ処理を行う場合と異なり，住所を含むページが含まれているとは限らない．スニペットを使用する場合も同様に住所が含まれることが少ない．そこで本研究では地物名一般に着目する．

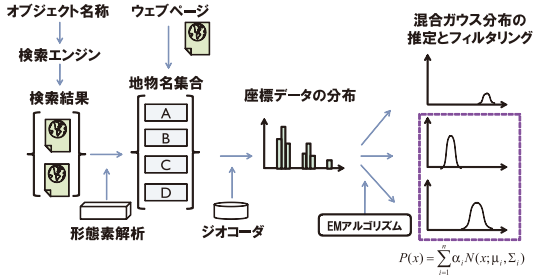

図 1　システムの構成
Fig. 1　System flow.

地物名一般を使用した場合の問題として，曖昧性による精度低下があげられる．例として，ページ内に「三越」という単語が含まれているだけでは，それがどの都市に存在する三越を表しているのかが明確でない．ページ中に含まれる建物名は位置の特定に役に立つにもかかわらず，曖昧性によって利用が困難になっている．

本研究では確率モデルの導入によって地物名の曖昧性に起因するノイズを取り除き，地域性の推定を行う．具体的な手法に関しては後述する．これによって「三越」といった曖昧性の高い語も，その他の地物名と組み合わせることで活用できる．確率モデルはより多くのデータが与えられるほど精度が高くなるという性質を持つため，多数のページを用いてオブジェクトの地域性を推定するという本研究における課題に適した方法であるといえる．取得するページ集合のサイズを大きくとることにより，地域性推定の精度は向上していくことが期待される．

一方，あまりにも多数の地物名を使用した場合，計算時間が増大化するため，地物名を出現頻度で降順にソートし，上位 m 件を使用するという方式をとる．また，高頻度の地物名が低頻度の地物名よりも優先して利用される仕組みとして，地物名の頻度が後述の EM アルゴリズムで使用される事象数に反映されるようにする．例として，地物名 A が地物名 B より 10 倍の頻度で現れたとすれば，地物名 A に対応する座標データが 10 倍の頻度で生起したと見なされ，混合ガウス分布の推定に反映される．ただし，事象数 $\eta(w)$ は式 (3) によって定義する．γ は重み，$tf(w)$ は地物名 w の出現頻度（text frequency），$cand(w)$ は地物名 w に対応する座標データの数（candidates），すなわち曖昧性を表す．事象数は整数でなくてはならないため，切り上げを行っている．

$$\eta(w) = \lceil \gamma \cdot tf(w)/cand(w) \rceil \tag{3}$$

重み γ を掛ける理由は，出現頻度が少なく曖昧性が高い場合にも，情報が失われないようにするためである．地物名 w に対応する座標データそれぞれについて $\eta(w)$ 個の事象が発生したと考え，EM アルゴリズムを用いて母分布の推定が行われる．

4.3 EM アルゴリズムを用いたモデル推定

地域性を表す確率密度関数の推定は EM アルゴリズムを用いて行う．EM アルゴリズムはパラメータの最尤推定量を求めるための計算手法であり，広く用いられている[16]．混合ガウス分布を EM アルゴリズムで推定する場合，以下のステップを繰り返して求める．ただし，x_k は事象，m は標本のサイズ，α_i は多次元ガウス分布 i に対する重み，ϕ_i は多次元ガウス分布 i のパラメータ μ_i と Σ_i をまとめたもの，Φ は混合ガウス分布の全パラメータをまとめたものを表す．また，プライム記号（′）を付けて表したパラメータ（α'，μ'，ϕ'，Φ'）は前ステップにおける値を表す．

$$\alpha_i = \frac{1}{m}\sum_{k=1}^{m}\frac{\alpha_i' p_i(x_k|\phi_i')}{p(x_k|\Phi')} \tag{4}$$

$$\psi_{ik} = \frac{\alpha_i' p_i(x_k|\phi_i')}{p(x_k|\Phi')}, \quad \Psi_i = \sum_{k=1}^{m}\psi_{ik} \tag{5}$$

$$\mu_i = \frac{1}{\Psi_i}\sum_{k=1}^{m}\psi_{ik}x_k \tag{6}$$

$$\Sigma_i = \frac{1}{\Psi_i}\sum_{k=1}^{m}\psi_{ik}(x_k-\mu_i)(x_k-\mu_i)^T \tag{7}$$

本研究においては，地物名に対応する座標データが事象 x_k の値となる．前節で述べたように，1 つの座標データに対し，$\eta(w)$ 個の事象を用意する．すなわち，1 つの座標データから発生する事象 $x_k, \ldots, x_{k+h}$ は同じ値を持つが，事象としては異なるものと見なされる．

EM アルゴリズムによって得られる結果は局所最適解であるため，疑似乱数を用いて初期値を繰り返し生成し，最も尤度の高いモデルを選択する．

4.4 フィルタリング

座標データから得られた各事象 x_k に対し，混合ガウス分布を構成する多次元ガウス分布 $N(x;\mu_i,\Sigma_i)$ のうち，x_k の生起確率が最も高くなる添え字 i' を割り当てることで，事象を分割することができる．

$$x_k \in c_{i'} \longleftrightarrow i' = \arg\max_i N(x_k;\mu_i,\Sigma_i) \tag{8}$$

このようにして得られた事象の集合 c_i を**クラスタ**と呼ぶ．ウェブページ中に含まれる地物名から得られたクラスタ集合において，要素数の少ないクラスタは地物名の曖昧性に由来するノイズである可能性が高い．そこで，これらのノイズを取り除くため，クラスタの要素数の降順でソーティングを行い，全事象 X のうち，割合 r 以上を含むよう，上位 ξ 件のクラスタだけを残す（式 (9)）．X は全事象の集合を表し，c_i はクラスタである．クラスタは要素数で降順にソートされ，$|c_i| \geq |c_{i+1}|$ を満たすように添え字を与えられている．フィルタリング後のクラスタ集合を K' で表す．

$$K' = \left\{ c_j \left| \sum_{i=1}^{\xi}|c_i| \leq r|X| \wedge r|X| < \sum_{i=1}^{\xi+1}|c_i| \wedge j \leq \xi \right. \right\} \tag{9}$$

最後に，K' に含まれるクラスタ c_j に対応するガウス分布 $N(x;\mu_j,\Sigma_j)$ を重み α_j で重ね合わせ，関数 $F(x)$ を求める．

$$F(x) = \sum_{c_j \in K'}\alpha_j N(x;\mu_j,\Sigma_j) \tag{10}$$

さらに，全体で積分したときに 1 となるよう，$F(x)$ の正規化を行う．これによって得られる確率密度関数 $P(x)$ を本研究における地域性のモデルとする．

$$\begin{aligned} P(x) &= \frac{F(x)}{\int_{R^2}F(x)dx} \\ &= \sum_{c_j \in K'}\frac{\alpha_j}{\sum_{c_u \in K'}\alpha_u}N(x;\mu_j,\Sigma_j) \end{aligned} \tag{11}$$

さらに，各座標データに対し，その座標において最も高い確率密度を持つガウス分布への割当てを行うことで，座標データのクラスタを生成する．これらの座標データのもととなった地物名を求め，重複を取り除くことで，各クラスタに対応する地物名集合を得る．視覚化の段階において，これらの地物名集合はページ/オブジェクトが関連する地物名の集合としてユーザに提示される．

5. 実　　装

本研究で提案された手法に基づき，ユーザが与えた任意のオブジェクト名称に対し，実時間でその地域性を視覚化するウェブアプリケーション「Location Classifier」を開発した．実装はサーバサイドで Perl，クライアントサイドで JavaScript を使用する AJAX の構成を用いた．また，推定された確率モデルの等確率密度線を描画するための地図インタフェースとして，Google Maps API を使用した[20]．

5.1 ページ収集

ページ収集のステップにおいては Yahoo! Web API を利用し，ユーザが入力したキー

ワード型検索クエリに対して検索結果を求める[19]．クエリが複数の単語から構成される場合，検索エンジンはその両者に関連するページ集合を取得する．検索結果の利用方法としては，1) ページ要約（スニペット）を利用する方法，2) ページ本体にアクセスして文書全体を利用する方法，の 2 通りの手法を用意している．短い応答時間を求める場合には要約，精度の高い結果を求める場合にページ本体の使用が望ましいため，ユーザが利用時に切り替えられるようにしている．利用する検索結果数はユーザが指定することができる．これによって応答までの時間を調節できる．

5.2　地物名抽出

地物名の抽出は形態素解析器 MeCab を用いて行う[18]．文書中より，「名詞，固有名詞，地域」の品詞を持つ形態素を収集する．あらかじめリストを与えておくことにより国名を取り除き，さらに，カタカナだけを含む地物名も取り除く．MeCab は辞書ではなく文脈を考慮して品詞の判定を行っているため，海外の地物名も取得してしまう．これらがジオコーダに与えられた場合，店舗名等とマッチしてしまうことがあるため，この処理を行っておく必要がある．

さらに，文書中に現れる郵便番号を取得して使用する．形態素解析結果から郵便番号の形式を持つ文字列（ハイフンで区切られた数字 3 桁と数字 4 桁）を抽出し，日本郵便[26] が提供している対応表を用い，住所への変換を行った．3 桁-4 桁という数字パターンは郵便番号以外でも使用される可能性があるため，若干のノイズも含まれるが，同一のパターンが複数回現れることは少ないと考えられるため，集約と出現頻度によるソーティングによって取り除くことができると考えた．

得られた結果に対して集約を行い，頻度の高い m 件の地物名のみを使用する．今回の実装では $m = 100$ を使用した．文書中での地物名の頻度 $tf(w)$ は EM アルゴリズムで事象数 $\eta(w)$ を求める際に使用するため，残しておく．

ページ集合から上位 m 件の地物名を取得する際，個々のページのサイズや検索エンジンの結果におけるランキングを使うといった手法も考えられるが，本研究では使用していない．

5.3　ジオコーディング

ジオコーディングの第 1 段階として，国土地理院が提供している国土数値情報のデータを使用し，都道府県名/市町村名の座標へのマッピングを行う[25]．これらのデータは無償で利用可能である．「公共施設」のうち，都道府県庁と市区役所・町村役場に対応する点データを使用した．市区町村名には曖昧性が存在しうるが，前述のとおり，複数の候補が存在する場合，文書中での出現頻度を候補の数で割ることにより，事象数 $\eta(w)$ を下げることで対応する．

第 2 段階として，Yahoo! が提供するローカルサーチ API を使用してより細かな地物名やランドマーク的地物名の取得を行った[19]．このデータを併用する理由は，国土数値情報では一般に広く知られた地域名や建物名等の情報が十分に提供されていないためである．

一般的なジオコーディングシステムと同様，Yahoo! ローカルサーチ API も 1 つの地物名に対して複数の座標を返すため，式 (3) の形で事象数 $\eta(w)$ を求める．

有償のジオコーディングサービスも存在するが，本実験では無償のデータを利用しても一定の精度を持つシステムを実装できることを示す．今後，国外にまでシステムの対象範囲を拡大していくためにも，取得が容易なデータを用いてもシステムが構築できることを目標とした．

5.4　EM アルゴリズム

座標データの集合に対し，EM アルゴリズムを適用して混合ガウス分布の推定を行う．EM アルゴリズムの計算には Perl を用いて独自に実装したプログラムを使用した．

式 (3) における γ（出現頻度を事象数に変換する際の重み）は初期設定で 2.0，しかしユーザが実行時に変更できるようにした．

5.5　フィルタリング

フィルタリングのステップにおいて，地物名の曖昧性に起因するノイズを取り除く．クラスタを要素数で降順にソートし，標本全体のうち割合 r が少なくとも含まれるように，上位のクラスタのみを残す．今回の実装では $r = 1/2$ を用いた．

5.6　確率密度関数による視覚化

ガウス分布は式 (2) によって表されるため，c を任意の正の数としたとき，等確率密度線は式 (12) における x の軌跡である．

$$(x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) = c \tag{12}$$

この軌跡は Σ^{-1} の最大固有ベクトル方向に長軸，最小固有ベクトル方向に短軸を持つ楕円となる．

実装されたシステムでは上位のクラスタに対応する楕円を実時間で描画し，オブジェクトの関連地域の視覚化を行っている．表示例を**図 2** に示した．大きさによってクラスタ順位を示すマーカと，等確率密度線を表す楕円とがともに地図上に描かれる．オブジェクト名称「うどん」で地域性の推定を行った場合，香川県と大阪府に上位クラスタが生成されている．

地図をズームインし，絞り込み検索を行うことで，特定の範囲においてオブジェクトが関連する地域を検索することも可能である．**図 3** では香川県内において特に「うどん」との

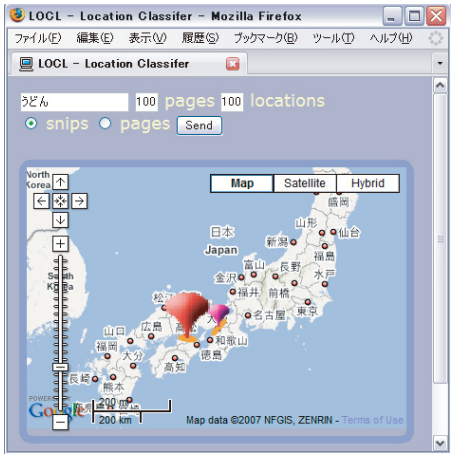

図 2　地図インタフェースにおける等確率密度線の描画
Fig. 2　Visualization of equal probability density contour.

図 3　対象地域を限定した推定
Fig. 3　Geographic relevance in a small area.

関連が深い地域を求めている．

ユーザがクエリに広域の地域名「香川」を加えることで，検索エンジンに送られるクエリが「うどん 香川」となり，香川県内のうどんに関するページ/スニペットが多く取得される．これらのページに含まれる地物名をジオコーディングし，地図に表示中の範囲内の座標のみを用いて推定を行うことで，詳細な関連地域を同定することができる．この際，対象範囲の面積が一定の閾値以下の場合，都道府県名は地物名として使用しない．実装では閾値として 10 万 km^2 を使用した．さらに，クエリ中に含まれる地物名（例の場合「香川」）もジオコーディングの対象にしないことによって，1 地点への集中を防いでいる．

6. 評　　価

実装されたシステムの評価を行うため，いくつかのカテゴリ（ジャンル）を選択し，そのカテゴリに属する多数のウェブページ/オブジェクトに対して評価実験を行った．これはウェブページ/オブジェクトの選択において恣意性が入ることを防ぐためである．

精度の評価指標として，要素数が最大となるクラスタの中心座標（μ）から正解と見なされる座標との距離，ならびに計算時間等を求めた．

ウェブページに対する評価としては，与えられたカテゴリに属するページを多数用意し，それぞれに対して地域性推定を行い，精度や計算時間の平均値を求めた．オブジェクトに対する評価としては，与えられたカテゴリに属するオブジェクト名をクエリとして順にウェブ検索を行い，得られた検索結果上位 100 件の要約（スニペット）を用いて地域性推定を行った．カテゴリとしては，以下を使用した．

ウェブページの地域性推定

- 各都道府県に関する観光情報サイトのトップ（47 件）
- FM ラジオ局のトップページ（47 件）
- 各都道府県に所在する商店街のページのトップ（47 件）

オブジェクトの地域性推定

- 土産物（49 件）
- 郷土料理（47 件）
- 祭り（34 件）

オブジェクトのリストを作成する際のソースとして使用したのは，Wikipedia における一覧表形式のコンテンツである[21]．これらの記事においては，カテゴリに属するオブジェクトの例がその関連地域とともにリスト形式で示されている．データ例を以下に示す．

【土産物】		【郷土料理】	
白い恋人	北海道	いももち	北海道
からし明太子	福岡県	イカのポンポン焼き	青森県
萩の月	宮城県	石焼き鍋	秋田県
八ツ橋	京都府	いものこ汁	岩手県
……		……	

実験に使用された計算機の CPU は Intel Xeon 3.20 GHz，RAM は 8 GB，OS は CentOS 5.0 である．

6.1　地物名と座標の取得結果

前節であげた各種カテゴリに対し，ウェブページならびにスニペット集合に形態素解析を適用し，地物名数に関する情報を得た．その結果，ページ 1 件あたり平均して 9.2 個の形態素が地物名（「名詞，固有名詞，地域」）として認識された．一方，スニペット 1 件あたり平均して 0.53 個の形態素が地物名と認識された．

表 1 の第 3 列ではページ 1 件あるいはスニペット 100 件から取得された地物名の種類数，第 4 列ではそれらに対して得られた座標データの種類数を示している．特にページの場合，カテゴリによって含まれる地物名の種類数が大きく異なることが読み取れる．さらに，1 つの地物名に対して多数の座標データが対応していることが示されている．

第 5 列以降では各カテゴリについて，上位のクラスタに含まれる地物名の種類数，ならびにそれに対応する座標データの種類数の平均をまとめてある．

表 1　全体と上位クラスタにおける地物名/座標データの種類数
Table 1　Amounts of place names and coordinates in top clusters and in total.

カテゴリ	件数	P_{total}	C_{total}	P_1	C_1	P_2	C_2	P_3	C_3
Pages 1	47	19.4	109.6	4.18	10.47	2.13	7.76	2.14	5.79
Pages 2	47	3.0	14.7	2.71	9.24	1.74	5.29	1.82	4.10
Pages 3	47	5.1	29.1	2.38	7.30	1.73	6.43	1.77	3.59
Obj. 1	49	46.2	220.5	3.14	6.32	3.32	7.81	3.20	5.91
Obj. 2	47	54.4	252.6	3.09	6.63	3.36	6.26	3.12	6.88
Obj. 3	34	59.1	299.6	2.13	3.41	1.72	3.41	1.87	4.25

P_{total}：全体における地物名の種類数，C_{total}：全体における座標データの種類数，P_j：第 j 位のクラスタに含まれる地物名の種類数，C_j：第 j 位のクラスタに含まれる座標データの種類数，Pages 1：観光情報サイト，Pages 2：FM 局のサイト，Pages 3：商店街サイト，Obj. 1：土産物，Obj. 2：郷土料理，Obj. 3：祭り（オブジェクトに関するデータは検索結果要約の上位 100 件から取得）

なお，上位クラスタほど多くの座標データを含むとは限らないのは，式 (3) で示したように，各座標データに対して重み γ と地物名の頻度（$tf(w)$）を掛け，曖昧性（$cand(w)$）で割って得られた回数分，事象として使用しているからである．

本節で対象としている全国的なスケールで推定を行った場合，最大クラスタに含まれる地物名の種類数は少なくなりがちである．これは広域的な地物名（都道府県名や都市名）と詳細な地物名（地区名，街区名等）の間で出現頻度に大きな差があり，これらを頻度順でソートして上位の地物名のみを使用するステップの存在により，詳細な地物名が取り除かれてしまうためである．結果として都道府県名や都市名のみを含む最大クラスタが生成されることが多い．

一方，6.5 節で述べる対象範囲を限定した推定の場合，現れる地物名は多様化し，クラスタにはより多くの種類の地物名が含まれ，対応するガウス分布も広がった分布を持つ傾向がある．

6.2　クラスタを構成する地物名集合

クラスタを構成する地物名の集合の具体例を**表 2** に示した．中括弧の中に含まれる地物名が 1 つのクラスタを構成している．クラスタはそれに含まれる事象数によってソートされているため，地物名の種類数の大小とは一致しない．地物名の後の括弧内で示されているのは地物名の出現頻度である．上位クラスタに含まれる地物名の多くはオブジェクトに対して直観的に関連の高い地物名であるという結果が得られた．同一の地物名が異なるクラスタに現れているのは，地物名の曖昧性から生じている．なお，本節以降で示される地物名において，「〜市」「〜町」等，地方自治体の種別が明示されているものは，郵便番号を通して得られた地物名を意味する．文書中から直接得られた地物名では地方自治体の種別は取り除かれている．

6.3　マッピングの精度

実験に使用したページ/オブジェクトに関しては，関連する都道府県名が文書中であらかじめ明示されているため，各都道府県の県庁所在地を正解座標としてマッピング精度の評価を行った．式 (3) における重みとして $\gamma = 2$ を使用した．クラスタを要素数の大きい順でソートし，第 j 位のクラスタの中心（μ_j）から正解座標への距離を求め，カテゴリごとに平均値を求めた（**表 3**）．

正解座標として県庁所在地を使用するという近似を行っているため，平均距離が 100 km 前後となることは比較的精度の高い結果であるといえる．特に，全国スケールの地図を表示させた状態で提示される結果としては，十分であると考えられる．5.6 節で述べたように，ユーザは全国スケールで地域性の推定を行った後，さらに詳細な情報が欲しい場合には，地図を特定の範囲にズームインしたうえで再推定させることができる．そのため，第 1 段階で

表 2「土産物」カテゴリにおける上位クラスタを構成する地物名集合
Table 2　Place names in the clusters for a “souvenir” category.

クエリ	上位クラスタを構成する地物名（クラスタの要素数の降順）
白い恋人	{ 札幌 (56), 北海道 (13)} { 北海道 (13)} { 石屋 (8)}
からし明太子	{ 福岡 (46), 博多 (13)} { 北九州 (8), 小倉北 (1), 北九州市 (2)} { 京都 (8), 北海道 (3), 川越 (2), 上尾 (2)}
萩の月	{ 仙台 (146), 宮城 (50)} { 東北 (18)} { 大河原 (14)}
菜の花まんじゅう	{ 箱根 (26), 小田原 (16), 箱根湯本 (2), 湯本 (1), 足柄下 (1)} { 東京 (18)} { 所沢 (16)}
いかなごくぎ煮	{ 神戸 (40), 兵庫 (36), 神戸市 (2), 宝塚 (2), 播州 (1)} { 明石 (26), 阿波 (2), 育波 (1)} { 瀬戸内 (8), 関西 (2), 瀬戸内海 (2)}
うなぎパイ	{ 浜松 (56)} { 静岡 (30)} { 大久保 (1), 水城 (1), 六甲 (1)}

表 3　マッピングの精度
Table 3　Mapping accuracy.

カテゴリ	件数	c_1	c_2	c_3	c_4
Pages 1	47	129.3	254.2	396.2	340.9
Pages 2	47	126.6	280.4	368.6	607.2
Pages 3	47	40.5	308.5	420.4	497.9
Obj. 1	49	72.5	254.6	343.1	387.0
Obj. 2	47	114.0	301.2	314.9	384.2
Obj. 3	34	71.3	211.4	235.6	344.3

c_j：第 j 位のクラスタ，Pages 1：観光情報サイト，Pages 2：FM 局のサイト，Pages 3：商店街サイト，Obj. 1：土産物，Obj. 2：郷土料理，Obj. 3：祭り（単位は km）

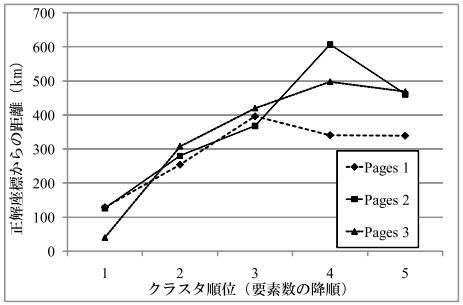

図 4　ページマッピングの精度
Fig. 4　Mapping accuracy for pages.

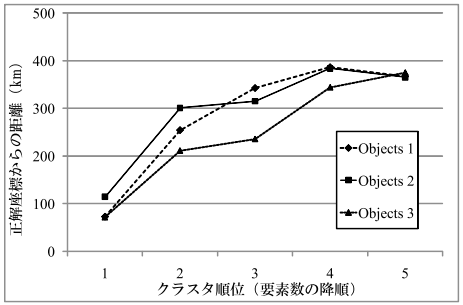

図 5　オブジェクトマッピングの精度
Fig. 5　Mapping accuracy for objects.

は都道府県レベルで正しくマッピングが行えればよい．

さらに，要素数の多いクラスタ（上位のクラスタ）ほど正解座標から近いという傾向が見られることにより，小さなクラスタをフィルタリングする手法の有効性が示されている．

図 4 および**図 5** ではページとオブジェクトを対象とした実験において，それぞれクラスタの順位ごとに，正解座標からの距離の平均値をグラフで示した．いずれのカテゴリにおいても要素数の大きいクラスタほど正解座標の近くにマッピングされている傾向が読み取れる．

6.4　複数地域へのマッピング

地域性のモデルとして混合ガウス分布を使用するメリットの 1 つとして，関連する地域が複数以上存在する場合，それぞれについてクラスタが取得されることがあげられる．複数の地域に関連するクエリで実験を行った例を**表 4** に示した．実験ではスニペット 100 件を用い，使用する地物名数として $m = 100$，式 (3) における重みとして $\gamma = 2$ を用いた．

「うどん」に関しては香川県を中心としたもの，関西を中心にしたもの，関東を中心にし

表 4　複数の関連地域へのマッピング結果
Table 4　Mapping to multiple relevant regions.

クエリ	上位クラスタを構成する地物名（クラスタの要素数の降順）
うどん	{ 香川 (32), 讃岐 (9), 多度津 (6), 坂出 (2), 琴平 (2), 善通寺 (2)} { 讃岐 (9), 大阪 (8), 滋賀 (6), 京都 (4), 観音寺 (2), 関西 (1)} { 東京 (10), 埼玉 (6), 所沢 (4), 松戸 (4), 成田 (2), 日野 (2), 武蔵野 (2)}
銘菓ひよ子	{ 東京 (170), 千住 (1), 台東 (1)} { 福岡 (114)}
お茶の産地	{ 静岡 (130), 静岡市 (2)} { 鹿児島 (18), 知覧 (10), さつま (2)} { 京都 (16), 篠山 (4), 丹波 (4), 滋賀 (2), 茶 (1)} { 三重 (26)}
りんごの産地	{ 青森 (112)} { 長野 (66)}
リゾート地	{ 沖縄 (38), 宜野湾 (2)} { 軽井沢 (28), 長野 (6)} { 箱根 (4), 東伊豆 (4), 横須賀 (2), 熱海 (2), 富士 (2), 房総 (1), 猿島 (1), 中伊豆 (1)}

表 5　対象範囲を限定したマッピングの結果
Table 5　Mapping within a specific area.

クエリ	対象範囲（南西隅と北東隅の経緯度）
上位クラスタを構成する地物名（クラスタの要素数の降順）	
うどん 香川	(133.4984, 33.8902) - (134.5884, 34.7902)
{ 丸亀 (20), 讃岐 (19), 坂出 (8), 多度津 (6), 宇多津 (2), 坂出市 (2), 金倉 (1), 土器町東 (1), 綾歌 (1)} { 讃岐 (19)} { 高松 (32), 高松市 (2), 鶴市 (1)}	
りんご 青森	(139.8532, 40.3591) - (141.7758, 41.6031)
{ 弘前 (34)} { 板柳 (6), 五所川原 (2), 大川 (2), 花岡 (1), 北津軽 (1)} { 末広 (2), 新鍛冶 (1), 下湯口 (1)}	
ピーナッツ 千葉	(139.8106, 34.9456) - (140.7719, 35.9524)
{ 八街 (28), 成田 (18)} { 館山 (6), 館山市 (4), 下真倉 (1)} { 佐倉 (10), 稲毛 (1)} { 新千葉 (1), 中里 (1), 稲毛 (1)}	
みたらし団子 京都	(135.6993, 34.9576) - (135.8195, 35.0560)
{ 下鴨松ノ木 (1), 下鴨宮崎 (1), 上京 (1), 下鴨本 (1), 下鴨 (1)} { 三軒 (2), 烏丸 (1), 上京 (1), 今出川 (1)}	
とんこつラーメン 福岡	(130.3440, 33.5674) - (130.4470, 33.6532)
{ 博多 (17)} { 長浜 (18), 博多 (17)} { 馬出 (2), 箱崎 (1), 吉塚 (1)}	

たもの等，複数の地域に対してクラスタが得られている．「銘菓ひよ子」は福岡と東京の土産として著名である．実験結果においても，東京と福岡に上位のクラスタが生成されている．また，「お茶の産地」の結果ではいずれも代表的な茶生産地域が抽出されている．「りんごの産地」の結果では青森と長野という主要産地が取得されている．「リゾート地」に関しては沖縄，長野に並んで，首都圏近郊のリゾート地の集合が 1 つの関連地域として抽出されている．

これらは全国的なスケールで推定を行った場合の結果であるが，より詳細な地域においても複数地点へのマッピングが行われることを次節で示す．

6.5　対象範囲を限定した推定

ユーザがオブジェクトに関する詳細な関連地域を知りたい場合，地図の対象範囲を狭め，その中に含まれる地物名のみを用いて推定を行うことができる（5.6 節）．このような利用方法の結果を**表 5** に示した．実験はスニペット 100 件，$m = 100$，ならびに $\gamma = 2$ という設定で行った．

限定された対象範囲の中で特にオブジェクトと関連の強い地域が取得されていることが示されている．一方，表 5 の結果は混合ガウス分布の利用における課題も示している．例として「りんご 青森」の結果において，要素数の降順で第 3 位のクラスタを構成する「末広」「新鍛冶」「下湯口」はいずれも弘前市内の地物名であり，本来ならば第 1 位のクラスタ（「弘前」）とまとめられることが望ましい．しかし，青森におけるりんごの代表的生産地である「弘前」の頻度が相対的に非常に高いため，1 地点への強い集中と見なされ，周辺の地

表 6　計算時間（単位：秒）
Table 6　Computation time (seconds).

カテゴリ	件数	Search	Fetch	Morph.	Geocode	EM	Total
Pages 1	47	0.010	0.775	0.066	1.848	1.499	4.197
Pages 2	47	0.011	0.380	0.028	0.410	0.143	0.972
Pages 3	47	0.012	0.513	0.040	0.778	0.391	1.734
Obj. 1	49	3.125	0.001	0.616	3.784	9.894	17.420
Obj. 2	47	4.362	0.001	0.574	4.521	9.638	19.097
Obj. 3	34	3.146	0.001	0.666	5.128	20.660	29.601

Search：ウェブ検索，Fetch：ページ取得，Morph.：形態素解析，Geocode：ジオコーディング，EM：EM アルゴリズム，Pages 1：観光情報サイト，Pages 2：FM 局のサイト，Pages 3：商店街サイト，Obj. 1：土産物，Obj. 2：郷土料理，Obj. 3：祭り

物名の分布と切り離され，別個のガウス分布でモデル化されてしまっている．これは混合ガウス分布に基づくモデルがガウス分布間の重なり合いを許容するために生じている現象である．このような状況は「ある県全体と関連するが，その中の特定の市と強く関連する」という状況と同じ構造をしているために，重なり合いを許容するべき場合とそうでない場合との間で切り分けが難しい．行政区域としての大きさ（「弘前」がどれだけの広がりを持つか）に応じて，事象を 1 点ではなく広がりを持ったものと見なすという補正が考えられるが，広がりを表すパラメータをどのように設定するかにおいて恣意性が入ってしまうという問題がある．「とんこつラーメン 福岡」においても，「博多」の出現頻度が高く，「長浜」に比べて曖昧性も低いために，それ自体でクラスタになってしまうという状況が生じている．「うどん 香川」「ピーナッツ 千葉」「みたらし団子 京都」のように，地物名の出現頻度の差が相対的に小さい場合には，近接する地物名どうしが 1 つのクラスタにまとめられ，関連地域が分散する傾向がある．

6.6　計算時間の評価

6.3 節で述べた実験において，それぞれのステップにかかった時間を**表 6** に示した．

オブジェクトを対象とする場合，全体の時間の中で EM アルゴリズムにおいて特に多くの時間がかかっていることが読み取れる．使用する地物名数を下げることで若干精度は下がるが EM アルゴリズムの計算時間は低減させられるため，ユーザが実行時にその値を調整することができるような実装にしている．

7. おわりに

本研究ではウェブページ中に含まれる地物名を用いた地域性推定手法を提案し，それに基づくオブジェクトレベルローカルサーチの実装について述べた．

図 6　フレーズに関連する地理的範囲の取得
Fig. 6　Geographic region relevant to a phrase.

評価実験の結果，要素数の多いクラスタは正解位置近くにマッピングされるため，クラスタのフィルタリングによって地物名の曖昧性から生じるノイズを低減させ，適切な地域にマッピングできていることを示した．また，計算時間の評価を行い，実時間で動作することを示した．

今後の展開として，応用と手法の両方に関して発展が考えられる．

応用の例として，実装されたシステム「Location Classifier」における検索クエリはオブジェクトに限定されず，フレーズを入力することもできる．たとえば「富士山が見えた」といったクエリを入れることで，**図 6** に示しているように，フレーズに関連する地域を求めることができる．このような機能はたとえば旅行者が訪問先を検討する際に有効である．旅行者の行動に関して，特定の地点で人気のある体験をブログから抽出するシステムの研究を筆者らは行ってきたが[5]，本システムでは逆に体験を指定し，関連する地域を取得することが可能になる．

もう 1 つの応用として，個々のオブジェクトに対して関連地域を事前に取得しておくことで，特定の地域に関連するオブジェクトの集合を返すという構成がある．たとえばユーザが指定した地理的範囲に強く関連する特産品を検索することができる．また，個々のページを対象にするだけでなく，サイト単位で地域性推定を行い，特定の地域に関連の深いウェブサ

イトを取得するといった応用も可能である．

手法の発展としては，確率モデルとして混合ガウス分布を使用したが，今後，さらに複雑なモデルを用いて記述するという方向性が考えられる．また，地物名の曖昧性から生じるノイズをより効果的に削減するために，既存のフィルタリング手法を改良して適用することがあげられる．また，推定に使用するパラメータ m と γ に関して，最適な値を探索するという課題がある．

実装面においては，地域的範囲を絞り込んで推定を行う際，ユーザが手動で地域名をクエリに追加する必要があるが，地図に表示されている範囲からユーザが興味を持っている地域を自動推定する機能を付けることが課題の 1 つであると考えている．

謝辞　本研究は，文部科学省研究委託事業「知的資産の電子的な保存・活用を支援するソフトウェア技術基盤の構築」，異メディア・アーカイブの横断的検索・統合ソフトウェア開発（研究代表者：田中克己），文部科学省科学研究費補助金特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」における計画研究「情報爆発時代に対応するコンテンツ融合と操作環境融合に関する研究」および，計画研究「情報爆発に対応する新 IT 基盤研究支援プラットフォームの構築」（研究代表者：安達淳，Y00-01，課題番号：18049073），若手研究（B）「ウェブ活用のための情報統合による信頼性判断支援」（研究代表者：手塚太郎，課題番号：18700086）によるものです．ここに記して謝意を表します．

参 考 文 献

1) McCurley, K.S.: Geospatial mapping and navigation of the Web, *Proc. 10th International World Wide Web Conference*, Hong Kong, China, pp.221–229 (2001).
2) Gao, W., Lee, H.C. and Miao, Y.: Geographically focused collaborative crawling, *Proc. 15th International World Wide Web Conference*, Edingburgh, Scotland, pp.287–296 (2006).
3) Zhou, Y., Xie, X., Wang, C., Gong, Y. and Ma, W.Y.: Hybrid index structures for location-based web search, *Proc. 14th ACM International Conference on Information and Knowledge Management*, Bremen, Germany, pp.155–162 (2005).
4) Matsumoto, C., Ma, Q. and Tanaka, K.: Web information retrieval based on the localness degree, *Proc. 13th International Conference on Database and Expert Systems Applications*, Aix-en-Provence, France, pp.172–181 (2004).
5) Kurashima, T., Tezuka, T. and Tanaka, K.: Mining and visualizing local experiences from blog entries, *Proc. 17th International Conference on Database and Expert Systems Applications*, Krakow, Poland, pp.213–222 (2006).
6) Nie, Z., Zhang, Y., Wen, J.R. and Ma, W.Y.: Object-Level Ranking: Bringing order to web objects, *Proc. 14th International World Wide Web Conference*, Chiba, Japan, pp.567–574 (2005).
7) Nie, Z., Ma, Y., Shi, S., Wen, J.R. and Ma, W.Y.: Web object retrieval, *Proc. 16th International World Wide Web Conference*, Banff, Canada, pp.81–90 (2007).
8) Buyukkokten, O., Cho, J., Garcia-Molina, H., Gravano, L. and Shivakumar, N.: Exploiting geographical location information of Web pages, *Proc. ACM SIGMOD Workshop on the Web and Databases*, Philadelphia, Pennsylvania (1999).
9) 森本泰貴，藤本典幸，長屋　務，出原　博，萩原兼一：Web を対象としたロボット型住所関連情報検索システムの開発，電子情報通信学会論文誌，Vol.J90-D, No.2, pp.245–256 (2007).
10) 相良　毅，有川正俊，坂内正夫：ジオリファレンス情報を用いた空間情報抽出システム，情報処理学会論文誌：データベース，41/SIG6(TOD7), pp.69–80 (2000).
11) 相良　毅，松浦啓一，佐藤　聡，志村純子：曖昧な地名照合手法を用いた生物種標本の地図ブラウザ構築，日本データベース学会 Letters, Vol.1, No.1, pp.39–42 (2002).
12) 有川正俊：地理的な位置をキーとしたコンテンツの獲得・管理手法，IT の深化の基盤を拓く情報学研究平成 15 年度研究成果報告書 A02 (2003).
13) 石川佳治：曖昧な位置情報に基づく空間問合わせの手法，日本データベース学会 Letters, Vol.6, No.2, pp.49–52 (2007).
14) Schneider, M.: Geographic data modeling: Fuzzy topological predicates, their properties, and their integration into query languages, *Proc. 9th ACM International Symposium on Advances in Geographic Information Systems*, Atlanta, Georgia, pp.9–14 (2001).
15) Shi, W. and Liu, K.: A fuzzy topology for computing the interior, boundary, and exterior of spatial objects quantitatively in GIS, *Computers & Geosciences*, Vol.33, No.7 (2007).
16) Bilmes, J.A.: A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian Mixture and Hidden Markov Models, Technical Report, University of Berkeley, ICSI-TR-97-021 (1997).
17) van Rijsbergen, C.J.: *Information Retrieval*, 2nd Edition, Butterworth & Co Publishers Ltd. (1979).
18) MeCab. http://mecab.sourceforge.net/
19) Yahoo! API. http://developer.yahoo.co.jp/
20) Google Maps API. http://google.com/apis/maps/
21) http://wikipedia.org/
22) Google マップ．http://maps.google.co.jp/
23) Yahoo! 地図情報．http://map.yahoo.co.jp/
24) Live Search Maps. http://maps.live.com/
25) 国土数値情報．http://nlftp.mlit.go.jp/ksj/

26) 日本郵便．http://www.post.japanpost.jp/

(平成 19 年 12 月 20 日受付)
(平成 20 年 4 月 7 日採録)

(担当編集委員　池田 哲夫)

手塚　太郎（正会員）

立命館大学情報理工学部メディア情報学科講師．2005 年京都大学大学院情報学研究科博士後期課程修了．博士（情報学）．主に地域情報検索システム，検索システムにおける信頼性向上の研究に従事．電子情報通信学会，日本データベース学会各会員．

近藤　浩之

京都大学大学院情報学研究科社会情報学専攻修士課程．2007 年京都大学工学部情報学科卒業．日本データベース学会学生会員．

田中　克己（正会員）

京都大学大学院情報学研究科社会情報学専攻教授．1976 年京都大学大学院修士課程修了．博士（工学）．主にデータベース，マルチメディアコンテンツ処理の研究に従事．IEEE Computer Society，ACM，日本データベース学会等各会員．