地域的支持度に基づくウェブページの信頼性評価とオブジェクトレベル検索

近藤　浩之† 　手塚　太郎† 　田中　克己†

† 京都大学大学院情報学研究科社会情報学専攻
〒 606-8501 京都府京都市左京区吉田本町
E-mail: †{kondo,tezuka,tanaka}@dl.kuis.kyoto-u.ac.jp

あらまし　Web ページやページ中に出現する地物などのオブジェクトがどのような地域的範囲から支持されているかを表す地域的支持度を導入する．本研究では、ページへの支持をハイパーリンクによって近似し，その広がりの度合いを信頼性判断支援に使用する．また，オブジェクトに関する多数のページを集約することによって，地域的支持度に基づくオブジェクト検索を行う手法を提案し、その評価について述べる．
キーワード　地域的支持度, オブジェクト検索, 信頼性判断

Reliability Evaluation of Web Pages and Object-level Search Based on Regional Supportness

Hiroyuki KONDO†, Taro TEZUKA†, and Katsumi TANAKA†

† Department of Social Informatics, Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo, Kyoto 606-8501 Japan
E-mail: †{kondo,tezuka,tanaka}@dl.kuis.kyoto-u.ac.jp

Abstract　As an index of reliability evaluation of web pages, we introduce “regional supportness” which indicates how regional area a web page and web object is supported from. Web object is , for examle, local features in web page. Supportness to pages approximates hyperlink, and we utilize its spreading degree for support of judgement of reliability. Moreover, we describe a method of object-level search based on regional supportness and its evaluation.
Key words　Regional supportness, topic level search, judgement of reliability

1. はじめに

ウェブ上の情報は急速に増加を続けており，日常生活における情報の獲得手段として欠かすことのできないものになりつつある．しかし，ウェブページを閲覧しただけでは，ウェブ上に書かれている情報が正しいかどうか判断することは容易ではない．

ウェブページの重要度を測る指標として，PageRank [1] やハブオーソリティモデル [2] などが広く知られているが，これらは信頼性を評価するための尺度ではなく、ウェブにおける人気度の評価に近い. 本研究ではウェブページの信頼性を評価するための新しい指標として、支持の空間的分布に着目した手法を提案する．また、個々のページの信頼性評価だけでなく、任意のオブジェクトに関して集約された情報に対する信頼性評価に拡張する．

ウェブ上の情報は世界中からアクセス可能であるが，利用者には地域的な偏りが存在することが多い．ウェブページは様々な地域の著者によって書かれており，ページからページへのリンクを支持と近似することによってページに対して様々な地域から支持されていると考えることができる．そこで，ウェブページがどのような地域から支持されているかについて可視化を行い，支持の地域性に関する尺度として地域的支持度を提案する．リンク元の可視化と地域的支持度を用いることによって，信頼性評価を行う．

地域性に関する検索として，Google マップ(注1)や Yahoo Local(注2)，Ask Business 検索(注3)，MSN City Guides(注4)といったローカルウェブ検索は現在幅広く利用されている．しかし，それらのシステムで提供される機能は非常に限られており，地

(注1): Google Local http://local.google.com/
(注2): Yahoo! Local Maps, http://maps.yahoo.com/
(注3): Ask Jeeves Local, http://local.ask.com/
(注4): MSN City Guides, http://local.msn.com/

域的支持の度合いのような情報を直接的に得ることはできない．

ウェブでの地域性は様々な層から成り立っており，支持度は地域性の一つの層である．その他の層の例として，ウェブページのコンテンツの地域性，ウェブページを利用している人の地域性，ウェブページの著者の地域性が含まれている．現状のローカル検索ではこのような異なる地域性の層を区別することはできない．

我々は地域的支持度に基づくオブジェクト検索を行うにあたり，地域性の多様な側面を表現したモデルを提案する．我々の手法を使うことによって，支持度のようなウェブコンテンツそのものに言及されていない地域的な情報の検索を行うスキーマを定義できる．このような地域的な支持の情報は以下のようなものである．

- 地元の住民の中でもっとも人気のある店やレストラン
- スポーツチームの支持の分布．あるチームは地域的な支持しかないが，別のチームは世界中から支持されている
- 政治家が支持されている分布
- 地方の専門家しかしらない地域

既存のウェブ検索エンジンではローカル情報を得るための複雑なクエリを入力することはできない．このような要求にこたえるためにウェブページの多層な地域性について考慮しなければならない．

我々の目的は多層な地域性の層を考慮することによって，地域的な支持度によるオブジェクトレベルの検索を行うことである．我々は地域的支持度によるオブジェクトレベルの検索を行うためにローカルウェブ検索の拡張について議論する．

具体的には，地理空間上に広がる確率分布を用いて地域性を表現し，ウェブページやオブジェクトに空間的インデックスを与えることで，地域的範囲を特定した検索を可能にする．

論文は以下のような構成となっている．第 2 章では関連研究について議論する．第 3 章では検索ページの信頼性評価について述べる．第 4 章では地域的支持度に基づくオブジェクトレベル検索について述べる．第 5 章ではローカルウェブ検索の定式化についてのべる．第 6 章では多層の地域性を評価するための手法とシステムを述べる．第 7 章では評価結果を示す．第 8 章では結論について述べる．

2. 関 連 研 究

この章では，信頼性支援とローカルウェブ検索の二つに分類して関連研究について議論する．

2.1 信頼性支援

山本らによる研究ではウェブの検索結果を集約することによって抽出した知識に対して真偽判定の支援を行っている [3]．中村らはウェブに対する検索のアンケートを行い，ウェブ検索結果の信頼性支援のためのシステムを開発した [4]

2.2 ローカルウェブ検索アプリケーション

McCurley による研究はウェブページを IP アドレスや住所，電話番号，その他の情報を利用して空間的な位置にマッピングを行った [5]．地図インターフェースと一般的なウェブブラウザを組み合わせた特別なブラウザが実装されている．

ウェブからの地理的な知識マイニングでは，Buyukkokten らが様々な新聞サイトにリンクしていたサイトの位置の偏りを観測した [6]．彼らは New York Times や San Francisco Chronicle にリンクしているサイトの IP アドレスを比較することによって，New York Times にリンクしているサイトが最も広範囲に分布していることを発見した．Gao らは地理空間上でウェブページを平等にクローリングするために，ウェブページの位置情報を用いたシステムを実装した [7]．Mei らはブログ上でよく議論されている話題の地理的な分布を調べた [8]．手塚らはウェブから抽出された情報を地理情報システムと結合させることにより，ランドマーク性等の新たな情報を定量的に評価することを可能にした .[10]．Gravono らはウェブ検索エンジンに送られてくるクエリを地域的なものと世界的なものに分類するためにサポートベクトルマシン (SVM) を提案した [9]．

ローカルウェブ検索のモデルとして，Chen らは空間的なクエリの索引とウェブページ間を相互に計算することで定式化した [11]．彼らはウェブページがもつ地域性の多層性やオブジェクトの地域性について議論していなかった．Zhou らは地理空間にローカルウェブコンテンツマップを効果的に使うために R*木を含むインデックス構造を結合して議論した [12]．松本らはウェブページの地域的な情報をフィルタリングしたり発見するためにローカルフィルターを提案した [13]．地域性の尺度としてページ中に含まれる地名の Minimum Bounding Rectangle(MBR) を計算した．馬らは話題の普遍性と地域性について議論した [14]．後者は夏祭りやスポーツのイベントといった地域空間の様々な位置に出現する話題を指し示すために利用された．Yang と Claramunt はウェブコンテンツの物理的な近接とセマンティックな類似性にマルコフモデルを適用した [15]．

3. 検索結果ページの信頼性評価

ウェブ検索を行ったときに，検索結果のページに書かれている情報は信頼できるかどうか判断することは難しい．しかし，それぞれの検索結果ページは様々なページからリンクされている．リンク元の著者は様々な地域に点在しており，それぞれの著者がそのページを支持していると考えられる．つまり，検索結果ページはそれぞれ，異なった地域から支持されていると考えることができる．検索結果ページが実際にどのような地域から支持されているか可視化し，どのような地域から検索されているかといった地域的支持度を用いることによって，検索結果ページの信頼性支援ができると考えられる．

3.1 地域的支持度によるウェブページの評価

ウェブページに対する地域的支持度とはウェブページが別の地域と比べ，ある地域のウェブユーザから支持されていることを意味している．言い換えると，支持者の分布は均等ではなく偏っている．支持者の位置をとらえるために，ウェブページ間のハイパーリンクを用いた．ウェブページからほかのウェブページへのリンクはリンクされているページへの支持と考えることができる．これらのリンクはページに対する支持の指標として使うことができる．リンクしているページは様々な種類のジオコーディング手法に基づいて空間上にマッピングされる．

図 1　Zion 国立公園への支持の可視化

図 2　Blue Ridge 国立公園への支持の可視化

表 1　平均ログ距離の逆数によるランキング

上位 3 件		下位 3 件	
国立公園	σ	国立公園	σ
Hot Springs	64.8	Hawaii Volcanoes	25.8
Blue Ridge	63.7	Denali	30.0
Zion	59.8	Acadia	41.0

表 2　カルバックライブラー情報量によるランキング

上位 3 件		下位 3 件	
国立公園	σ	国立公園	σ
Hot Springs	71.1	Hawaii Volcanoes	30.0
Glacier	67.9	Everglades	33.6
Badlands	60.0	Big Bend	41.0

このように地域的支持度を考えることによって，あるページがどのような地域からリンクされているかを知ることができる．

図 1-2 はリンク元の IP アドレスに基づいた国立公園のウェブサイトに対する支持を可視化したものである．それぞれのサイトは異なるところからリンクされており，地理空間上での人気の分布を反映している．

地域的支持度を示す指標として，二つの尺度を導入する．1) リンク元ページと対象ページ間のログの平均距離の逆数 2) IP アドレスをランダム生成したときの一般的なホストの位置の分布とリンク元の位置の分布間でのカルバックライブラー情報量．前者はウェブサイト周辺でのリンク元の密集度を表しており，後者は対象サイトの位置に限らず，様々なエリアでの偏った分布を表している．ただし，位置情報は登録されているウェブのホストの住所から得ている．これらの値によるランキング結果は表 1-2 である．20 の国立公園を例に用いて，標準偏差を計算した．Hot Springs や Blue Ridge は近隣から高く支持されており，一方，Hawaii Volcanoes やアラスカにある Denali は遠くから支持されている．

このような手法を用いることによって，ウェブページがどのような地域から支持されているか可視化でき，支持されている地域を発見することができる．検索ページに対して，このような可視化を行うことによって，どのような地域から支持されているかわかり，信頼性評価の基準となる．

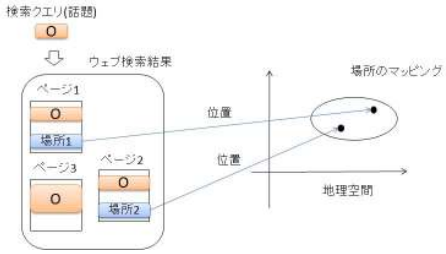

図 3　コンテンツの関連性によるウェブオブジェクト検索

次の章では，オブジェクトレベルでの地域的支持度とコンテンツの関連性を評価するために文章解析を利用した方法を提案する．

4. 地域的支持度のためのローカル検索の拡張

既存のローカルウェブ検索の特徴の一つとしてローカル情報の検索はページ単位であるということがあげられる．ページはウェブの情報の絶対的な単位ではない．ひとつのページは様々な地理的なオブジェクトの情報を含んでいる可能性があり，また様々なページに分割されていることも考えられる．

異なる特徴として，今日の多くのローカルウェブ検索エンジンはコンテンツベースのシステムであるという点がある．地理空間との関連性はページ内に何が書かれているかに基づいて決められる．しかしながら，ウェブオブジェクトがコンテンツの関連性の他にも持っている様々な地理的特徴がある．

この論文では，我々はこれらの観点からオブジェク検索を行うためにローカルウェブ検索を拡張するためのスキーマを提案する．

4.1　ウェブオブジェクト検索

ページはウェブページの情報の絶対的な単位ではない．一つのウェブページが様々なオブジェクトを含んでいたり，ひとつのオブジェクトが様々なウェブページ集合によって述べられていたりする可能性がある．必然的に，ページは地理上にマッピングされる絶対的な単位ではない．

我々はこのような考え方を表すためにウェブオブジェクトという言葉を使う [16] [17]．ウェブオブジェクトとはウェブページや複数のウェブページを集約した情報のことを指す．ウェブオブジェクトの例として 1) ウェブページ 2) ある主題を述べているページ集合 3) ある主題に関係するウェブページ上の情報がある．

この論文では，我々はウェブオブジェクトの二つの種類について議論する．ひとつはウェブページについてである．もうひとつは，様々なページに分布しているある主題に関する情報についてである．我々はそれをオブジェクトという．図 3 で示すように，ページとページ中で表現されているオブジェクトの両方が検索クエリの対象となる．ローカルウェブ検索はローカルウェブページ検索である必要はない．

ウェブオブジェクトにおける地域的支持度とはウェブオブジェクトが別の地域と比べ，ある地域のウェブユーザから支持され

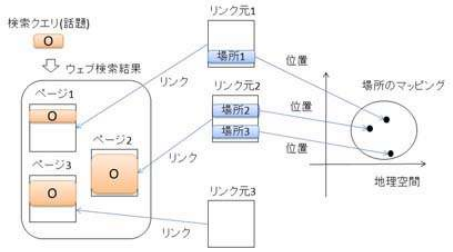

図 4　支持度によるウェブオブジェクト検索

ていることを意味している (図 4)．ページに対する地域的支持度をウェブオブジェクトに拡張している．

5. ローカル検索の定式化

この章では，我々は一階述語論理を用いたローカルウェブ検索の定式化を行う．この目的はこの論文で議論されている我々の新しいアプローチを拡張されたローカルウェブ検索に包含させるためである．このモデルを用いてオブジェクトレベルでのローカルウェブ検索について議論するために利用する．

5.1 地域性のあるクエリ

ローカルウェブ検索のクエリは述語 $L(x)$ によって表現され，ローカルウェブ検索システムは $L(x)$ が真のとき，ウェブオブジェクトとして検索される．

$$L(x) = P(x) \wedge R(x, \{U_\zeta\}, \{w_\xi\}) \tag{1}$$

x はウェブオブジェクトである．P は一般的な属性による述語を表し，R は空間的な関係を示している．

空間的な関係の定式化はローカルウェブ検索の能力をモデリングする上で重要な問題である．ウェブオブジェクトが空間的な位置とどの程度関連しているか考えなければならない．

異なるレベルでのウェブオブジェクトの空間関連性を表現するためにポテンシャル関数を用いる．ポテンシャル関数はユークリッド空間 R^d から $R^{d'}$ への関数である．我々は主に $d' = 1$ である場合のスカラーポテンシャルについて考える．

空間関係 R はポテンシャルを使って定義された様々な空間関連性 $r_{i,j}$ の標準結合で表わされる．

$$R(x, \{U_\zeta\}, \{w_\xi\}) = \bigwedge_{i=1}^{n} \bigvee_{j=1}^{m_i} r_{i,j}(x, \{U_\zeta\}, \{w_\xi\}, \{A_\phi\})$$

$r_{i,j}$ は，ユーザ要求ポテンシャル集合 $\{U_\zeta\}$，パラメータ集合 $\{w_\xi\}$，オブジェクト属性ポテンシャル集合 $\{A_\phi(x)\}$ を用いることによって定義される空間関連性である．

5.2 地域性の多層性

ユーザの要求やウェブオブジェクトには多層な地域性が存在し，先ほど述べたようなポテンシャル関数で表現することができる．以下に示すものは地域性の層の例である．

コンテンツの関連性の層は, ウェブオブジェクトがどうくらい地域と関連しているかを示している．

著者の層は，各地域でどれだけのウェブオブジェクトが書かれたりしているかを示している．

ユーザの層は，各地域でウェブオブジェクトがユーザからどの程度利用されているかを示している．

支持の層は，各地域でウェブオブジェクトがユーザによってどの程度支持されているかを示している．

我々はコンテンツ関連性ポテンシャルと支持ポテンシャルの評価について議論する．

5.3 オブジェクトの地域的属性

ウェブオブジェクトが持つ地域性はあいまいであることが多い．我々のモデルでは，ポテンシャル関数を用いることによって定式化する．

オブジェクト属性ポテンシャルとはウェブオブジェクトに付随しているポテンシャルのことである．検索結果で要求されている空間関係は U_ζ,A_ϕ という制約として表現される．ウェブオブジェクト x についてのオブジェクト属性のポテンシャル $A_\phi(x)$ はユークリッド空間 R^d から R への関数である．

$$A_\phi(x) : R^d \to R \tag{2}$$

ϕ は地域性の層を表している．典型的なウェブオブジェクト x は例えば，ベクトル $(A_{content}(x),\ A_{author}(x),\ A_{use}(x),\ A_{support}(x))$ のような複数の地域性の層を表している

オブジェクト属性ポテンシャルは関数の種類として表わされる．中心から遠ざかるほど減っていくポテンシャルを表現するために，混合ガウス分布を用いた．

5.4 オブジェクトの地域的属性の評価

ウェブマイニングによって得られた空間情報は，多くの場合分散しており店や地域の集まりとなっている．例えば，ウェブページ上に出現する地名はページと関連している連続的な地域を表しているものではなく，地理空間上では互いに離れたばらばらのエリアにマッピングされる．

我々はパラメトリックモデルを用いて，そのようなデータからオブジェクトの地域性を評価した．パラメトリックモデルでは，サンプルが混合ガウス分布のようなある確率モデルから生成されていると考える．パラメータを評価するために最大尤度法や EM アルゴリズムを用いた．この論文では，我々はこの手法を地域性のポテンシャル関数を評価するために用いる．

ウェブコンテンツのテキストデータから抽出したオブジェクト属性ポテンシャルを評価する手法を第 6 章で詳細に議論する．

6. 地域性の評価

この章では，ウェブページとウェブオブジェクトの抽出手法とコンテンツの関連性，地域的支持度の評価について述べる．コンテンツの関連性を取得するために，ウェブ文章のテキストを解析している．支持を取得するために，ハイパーリンクを用いた．それぞれの結果を地理空間上でのポテンシャル関数として表現し，それに対する評価を行う．

をオブジェクトに関連するページ集合とし，その集合に含まれる地名やそのページ集合にリンクしているページ集合をポテンシャル関数を計算するために用いる．

6.1 ソースデータ

ページにおける地域情報とは対象のページや対象ページにリ

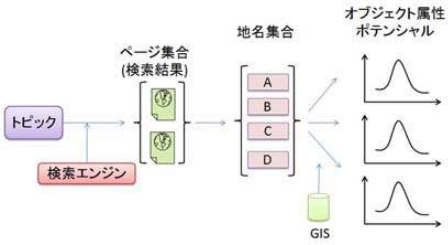

図 5　ウェブオブジェクトの地域性判断

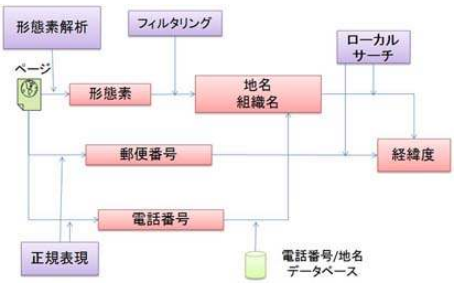

図 6　ページの解析とジオコーディング

ンクしているページ集合に含まれている地名，組織名，電話番号，郵便番号を指す．オブジェクトにおける地域情報とはオブジェクトを地域情報に関するクエリとして検索した時の検索結果集合やそれらのページにリンクしている地域情報を含むページ集合に含まれる情報を指す．

6.2　ウェブページの地域情報の抽出

ウェブページのコンテンツから形態素解析を用いて，地名・組織名を得て，ジオコーディングを行った．また，正規表現を用いて電話番号・郵便番号を得た (図 6)．

6.3　オブジェクトの地域情報の抽出

オブジェクトのキーワードをクエリとして検索エンジンに送り，検索結果ページ集合を得る．ページ集合をひとつの文章に統合し，地名のような地域情報を一つのページを対象としたときと同様の手法に抽出する．

図 3 のように，ウェブページはしばしば，さまざまなオブジェクトについて述べられており，ページによって出現するオブジェクトの重要性が異なる．しかし，ページ内でのオブジェクトの重要度を判断するための辞書を用いることを避けるため，すべてのページを同様に計算した．

6.4　ジオコーディング

我々はウェブサービス(注5)を用いて地名や組織名，郵便番号の経緯度を得た．

電話番号は総務省が公開している市外局番の対応表を用いて知名に変換した．さらに，地名をウェブサービスを用いて経緯度を得た．図 6 にページの解析システムの流れを示している．

6.5　コンテンツとの関連性の評価

ウェブオブジェクトに含まれている地名を用いることによって，コンテンツの関連性を評価する．

ウェブページの場合，そのページから地名が抽出され，ジオコーディングされて属性ポテンシャル関数に変換される．

(注5)：Yahooo!ローカルサーチ API, http://developer.yahoo.co.jp/map/

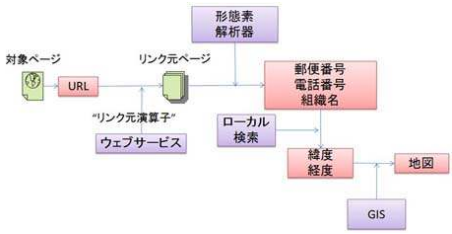

図 7　支持のポテンシャルの評価

オブジェクトの場合，オブジェクト名を検索エンジンのキーワードとして送り，検索結果ページ集合を得る．そのページ集合を単一のドキュメントに集約し，その中に含まれる地名をジオコーディングする．

6.6　地域的支持度の評価

我々は地域的支持度 (支持のポテンシャル) を評価するために，次のような手法を用いた．

我々はハイパーリンクをウェブコンテンツに対する支持の指標としている．もしページ A がページ B へリンクしているとしたとき，ページ A の著者はページ B のコンテンツを支持していると考えることができる．リンク元 (A) の地域情報を集めることによって，我々は B に対する支持のポテンシャル (地域的支持度) を評価する．

このシステムは対象ページにリンクしているページ (リンク元) を得るためにウェブ検索エンジンを利用する．リンク元を得るために我々のシステムは Yahoo!ウェブサービス API を用いた．リンク元のウェブページ集合から前述の手法で地名を抽出した．これらの地名から地域的支持度 (支持のポテンシャル) を得て，評価を行った．

オブジェクト単位の地域的支持度も同様に，検索結果集合にたいしても適応される (図 4)．コンテンツの関連性の評価と同じ手法で，ページ群は一つの文章に集約され，先ほど述べた手法によって地域情報が抽出される．図 7 は支持の解析システムの流れを示している．

6.7　EM アルゴリズムによるポテンシャルの評価

我々のシステムでは混合ガウス分布を学習することによってポテンシャルを評価している．このモデルは重み α_i で，$N(x;\mu_i,\Sigma_i)$ の様々なガウス分布の線形結合として表わされる．

$$
\begin{aligned}
P(\mathbf{x}) &= \sum_{i=1}^{n} \alpha_i N(x;\mu_i,\Sigma_i) \\
&= \sum_{i=1}^{n} \frac{\alpha_i}{2\pi\sqrt{|\Sigma_i|}} \exp\left(\frac{-(\mathbf{x}-\mu_i)^T\Sigma_i^{-1}(\mathbf{x}-\mu_i)}{2}\right)
\end{aligned}
$$

$$
\sum_{i=1}^{n} \alpha_i = 1 \qquad (3)
$$

パラメータ α_i，μ_i，Σ_i は EM アルゴリズム [18] を用いることによって評価される

我々の手法では，このモデルのガウス分布の数を赤池情報量基準 (AIC-Akaike Information Criterion-) [19] を利用することによって決めている．

$$A = -2\log f(\mathbf{x}|\mathbf{\Phi}) + 2N \tag{4}$$

A は AIC，$\mathbf{x}$ はサンプル，$\mathbf{\Phi}$ は評価されたパラメータのベクトル，$f(\mathbf{x}|\mathbf{\Phi})$ は尤度関数，N はパラメータ数である．数多くの試行を行った後，AIC がもっとも低いモデルを選ぶ．

結果として，我々は混合ガウス分布として表現されたウェブオブジェクトの地域性ポテンシャルを得ることができる．

6.8 枝 刈 り

ウェブページはコンテンツの中に様々な地名を含んでいる．それらの中には同じ単語でも地名や人名など両方の意味でとらえることのできる多義語やコンテンツとの関連性の低い単語が存在する．このようなノイズを取り除くために，関連性の低いガウス分布を取り除く．同じページ上に近隣の地名が出現しないウェブの文章が現れた場合それをノイズであるという仮定に基づいて，枝刈りを行う．

我々は重み α_i が最も大きい k 個のガウス分布を残し，これらのガウス分布 (クラスタ) に属しているページに出現している地名の割合 r を閾値とし，それよりも大きいものを選択した．

枝刈り後の分布は次の式であらわされる．

$$P(x) = \sum_{j \in J} \alpha_j N(x; \mu_j, \Sigma_j) \tag{5}$$

J は重み α_i が大きい上位 k 件のガウス分布の索引である ($|J| = k$)．

7. 実 験

我々は具体的な領域でウェブページとオブジェクトに対して，提案した手法の実験を行った．ウェブページの地域性の評価を行うために，日本の各都道府県から選択した 47 の地域ポータルサイトに対して実験を行った．地域的支持度を計算するために，我々はそれぞれのサイトのリンク元を 100 ページ収集した．オブジェクトについては，我々は日本の 31 のプロサッカーチームを用いた．我々の提案した手法を使うことによって，我々はコンテンツの関連性や地域的支持度を評価した．それぞれのチームで，チーム名をクエリとして既存の検索エンジンに入力することによって得られた 10 件のウェブページを使った．それぞれのページに対して，10 件のリンク元のを得ることによって，それぞれのチームに対して 100 件のリンク元を得た．

ガウス分布のポテンシャルは二つのパラメータ μ と Σ を持っている．その共分散行列 Σ はそのモデルがどのように分布しているかを示している．例えば，もし σ_x^2 が大きければ，そのサンプルが地図上に広く分布している．我々は分布のレベルを測るために，共分散行列 $||\Sigma||$ を用いる．$||\Sigma||$ の値はコンテンツの関連性や支持を反映しているサンプルがどのくらい空間的に広がって分布しているのかを示している．もし $||\Sigma||$ が大きければ，ウェブオブジェクトが地域的ではないと判断する．$||\Sigma||$ の小さいウェブオブジェクトを，地域性なウェブオブジェクトであると考える．

我々の実験では，行列のノルムを計算するためにスペクトルノルムを利用した．それは Σ の行列固有値のルート $\sqrt{\lambda_{max}}$ である．

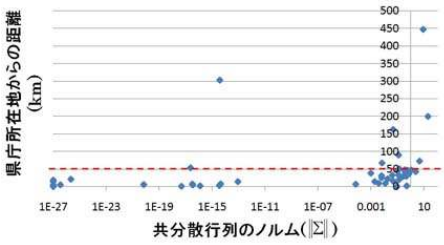

図 8 ポータルサイトのコンテンツと地域の関連性

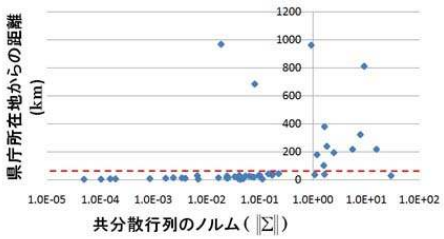

図 9 ポータルサイトの地域的支持度

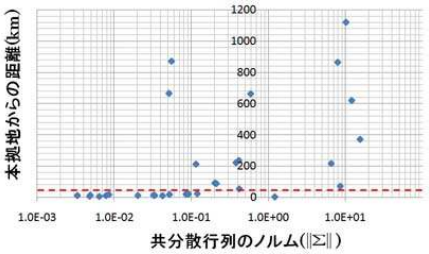

図 10 サッカーチームのコンテンツと地域の関連性

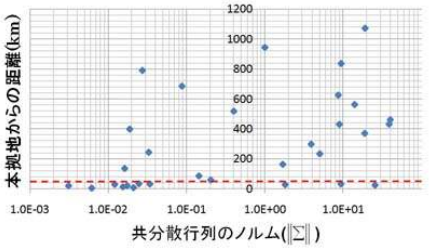

図 11 サッカーチームの地域的支持度

地域ポータルサイトについて，県庁所在地の座標をオブジェクトの物理的な位置として利用した．また，サッカーチームについては本拠地の座標を利用した．

7.1 地域性評価の実験

物理的な位置と重みが最大のガウス分布の平均ベクトル μ_i との二点間の距離 d を計算し，それを評価尺度として用いた．図 8-11 は結果をグラフにしたものである．Y 軸は物理的な位置と重みが最大のガウス分布の平均ベクトル μ_i との距離 d を示している．グラフは d と $||\Sigma||$ の関係を図示している．地域性の高いウェブオブジェクトの平均ベクトルは物理的な位置ととても近いことがわかる．X 軸は共分散行列 $||\Sigma||$ のスペクトルノルムである．Y 軸はオブジェクトの物理的な位置と評価された平均ベクトル μ_i 間の距離 d である．

共分散行列のノルムは様々な値をとる一方，物理的な位置と

表 3　コンテンツの関連性の地域性によるサッカーチームのランキング (上位 3 件と下位 3 件)

オブジェクト名	Σ のノルム	||Σ|| による順位	距離 (km)	距離による順位
ガンバ大阪	3.3E-03	1	10.4	6
FC 東京	4.8E-03	2	5.4	3
湘南ベルマーレ	4.9E-03	3	11.7	9
コンサドーレ札幌	1.0E+01	29	1120.8	31
モンテディオ山形	1.2E+01	30	619.2	26
大分トリニータ	1.5E+01	31	370.1	25

表 4　地域的支持度によるサッカーチームのランキング (上位 3 件と下位 3 件)

オブジェクト名	Σ のノルム	||Σ|| による順位	距離 (km)	距離による順位
東京ベルディ 1969	3.1E-03	1	18.7	4
横浜 F・マリノス	6.2E-03	2	3.5	1
柏レイソル	1.2E-02	3	28.0	8
京都サンガ	2.6E+01	29	23.4	6
横浜 FC	3.9E+01	30	430.6	22
サガン鳥栖	4.0E+01	31	460.3	23

平均ベクトル μ_i 間の距離はコンテンツの関連性ポテンシャルの場合小さくなり，地域的支持度の場合比較的分散することがわかった．

点線は距離が 50km であることを示している．ポータルサイトの支持のポテンシャルの場合 47 サイト中 35 サイトの距離が 50km 以内にあることがわかった．サッカーチームの支持のポテンシャルの場合，31 サイト中 11 サイトが 50km 以内にあることがわかった．サッカーチームの場合に比べ，ポータルサイトへの支持は物理的な位置の中心に近いことがわかる．

表 3-4 はそれぞれコンテンツ関連性や地域的支持度によってサッカーチームをランキングした結果である．我々は共分散行列を用いて地域性を定義した．ノルムが小さくなればなるほど，地域性が高い．中心付近での地域性ポテンシャルの強さを示している．表には，上位 3 件と下位 3 件の結果を記した．最も右の列は物理的な位置と評価された平均ベクトル μ 間の距離 d によって並べ替えた時の順位を示している．ノルム $||\Sigma||$ と距離 d の相関関係を示しているが，これらは全く同じではない．

表 5 はポータルサイトについて県庁所在地と評価された地域性 (μ_i) の中心との距離が比較的近いことを示している．サッカーチームについては本拠地と μ_i の間の距離がとても離れていることがわかる．しかし，支持のポテンシャルがコンテンツの関連性のポテンシャルより比較的幅広く分布していることがわかる．実際の実験では，システムはウェブオブジェクトを地域性が十分高い時だけ，地域的だと判断した．これは共分散行列のノルムを用いることによって計算できる．

7.2　検索アプリケーション

我々の提案した手法によって抽出された情報を利用したアプリケーションを描いた．図 12 はユーザが “横浜で支持されているサッカーチーム”とで検索した時の例である．そのシステム

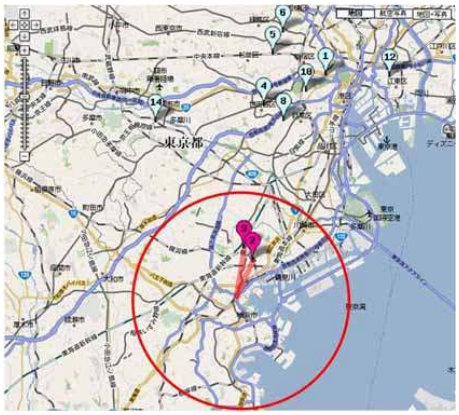

図 12　地域的支持度を用いたサッカーチームの検索例

表 6　“横浜で支持されているサッカーチーム”での検索結果

オブジェクト名	距離 (km)	||Σ|| による順位
横浜 F・マリノス	4.3	2
湘南ベルマーレ	5.3	9
川崎フロンターレ	19.9	8
FC 東京	21.2	4
バンフォーレ甲府	21.1	14

は Google Maps API(注6)をインターフェースとして実装した．

システムは日本でのサッカーチームへ支持しているポテンシャルを計算している．例えば横浜市といった地理的な位置にクエリがマップされると，チームはクエリの位置からもっとも大きいガウス分布の平均ベクトル μ_i でランキングされる．この場合，図 12 でピンクのマーカによって示されているようなもっとも近い 2 つのチーム (横浜 F・マリノスと湘南ベルマーレ) である．線はクエリの位置と近隣のマーカの間に引かれている．マーカーの番号は共分散行列のノルムによるランキング結果を示している．クエリの位置に近い平均ベクトルのチームは表 6 に記されている．

湘南ベルマーレは共分散行列のノルムが 9 位と比較的低いため，支持の分布が幅広く分散していてることがわかる．一方，横浜 F・マリノスは 2 位であるので集中した分布になっている．実際．横浜 F・マリノスは横浜市に位置しているサッカーチームだけで，湘南ベルマーレは横浜の南の海岸沿いに位置しているので，そのような結果は直観と一致している．

8.　結　　論

我々は検索ページの信頼性評価を行うために，ページに対する地域的支持度を提案した．さらに，我々はポテンシャル関数に基づいてローカルウェブ検索の定式化を導入し，コンテンツと地域的支持度を使うオブジェクトレベル検索についての手法と評価についてのべた．我々の実験では地域情報を伴うウェブページとオブジェクトが我々のシステムを使うことによって地理空間上に正確にマッピングされることを示した．

既存のローカルウェブ検索のアプリケーションは “地元の住

(注6): http://www.google.com/apis/maps/

表 5　コンテンツの関連性と支持による平均距離とノルム

ウェブオブジェクトの種類	コンテンツの関連性による平均距離 (km)	コンテンツ関連性による平均ノルム	支持の平均距離 (km)	支持の平均ノルム
ポータルサイト (48 サイト)	47.4	0.96	125.5	1.8
サッカーチーム (31 チーム)	211.1	2.1	308.4	6.7

民から支持されているレストランのページ”のようなクエリに答えることがでなかった．我々のモデルと実装したシステムは様々な課題を含めるために拡張されている．

提案した手法は様々な方法に適応することができる．ローカルウェブ検索はユーザによって与えられた地域的なクエリへのそれぞれのページの妥当性に基づいてウェブページをランクすることができる．一方，地域の知識抽出システムは地図インターフェース上で，与えられたキーワードの関連のある地域を可視化することができる．

我々の手法を用いることによって，地域性や地域的支持度によってページやオブジェクトを索引付けすることができる．ローカルウェブ検索を拡張することによって，既存のスキーマでは実現されなかった新しいオブジェクトレベルのローカルウェブ検索を実現できる．

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すものとします．

ここに記して謝意を表すものとします．

文　献

[1] Sergey Brin and Lawrence Page, “The Anatomy of a Large-Scale Hypertextual Web Search Engine,” Computer Networks and ISDN Systems, Vol.30，pp.1-7， 1998

[2] Jon M. Kleinberg, “Authoritative sources in a hyperlinked environment,” Journal of the ACM Vol.46, No5, pp.604-632 , 1999

[3] Yamamoto, Y., Tezuka, T., Adam, J. and Tanaka, K.: Honto? Search: Estimating Trustworthiness of Web Information by Search Results Aggregation and Temporal Analysis, Proceedings of a joint conference of the 9th Asia-Pacific Web Conference and the 8th International Conference on Web-Age Information Management, 253-264, HuangShan, China, 2007

[4] Nakamura, S., Konishi, S., Adam, J., Ohshima, H., Kondo, H., Tezuka, T., Oyama, S., Tanaka, K.: Trustworthiness Analysis of Web Search Results, Proceedings of the 11th European Conference on Research and Advanced technology for Digital Libraries, 38-49, Budapest, Hungary, 2007

[5] McCurley, K. S.: Geospatial mapping and navigation of the Web, Proceedings of the 10th International World Wide Web Conference, 221-229, Hong Kong, China, 2001.

[6] Buyukkokten, O., Cho, J., Garcia-Molina, H., Gravano, L. and Shivakumar, N.: Exploiting geographical location information of Web pages, Proceedings of the ACM SIGMOD Workshop on the Web and Databases, Philadelphia, Pennsylvania, 1999.

[7] Gao, W., Lee, H. C. and Miao, Y.: Geographically Focused Collaborative Crawling, Proceedings of the 15th International World Wide Web Conference, 287-296, 2006.

[8] Mei, Q., Liu, C., Su, H. and Zhai, C.: A Probabilistic Approach to Spatiotemporal Theme Pattern Mining on Weblogs, Proceedings of the 15th International World Wide Web Conference, 533-542, 2006.

[9] Gravano, L. , Hatzivassiloglou, V., and Lichtenstein, R.: Categorizing web queries according to geographical locality, Proceedings of the 12th International Conference on Information and Knowledge Management, pp. 325-333, New Orleans, Lousiana, 2003.

[10] Tezuka, T., Kurashima, T. and Tanaka, K.: Toward Tighter Integration of Web Search with a Geographic Information System, Proceedings of the 15th World Wide Web Conference, 277-286, Edinburgh, Scotland, 2006.

[11] Chen, Y., Suel, T., and Markowetz, A.: Efficient query processing in geographic web search engines, Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data, pp. 277-288, Chicago, Illinois, 2006.

[12] Zhou, Y. Xie, X., Wang, C., Gong, Y., Ma, W. Y.: Hybrid index structures for location-based web search, Proceedings of the 14th ACM International Conference on Information and Knowledge Management, pp. 155-162, Bremen, Germany, 2005.

[13] Matsumoto, C., Ma, Q. and Tanaka, K.: Web Information Retrieval Based on the Localness Degree, Proceedings of 13th International Conference on Database and Expert Systems Applications, pp. 172-181, Aix-en-Provence, France, 2004.

[14] Ma, Q., Matsumoto, C., and Tanaka, K.: A Localness-Filter for Searched Web Pages, Proceedings of the 5th Asia-Pacific Web Conference, pp. 525-536, Xian, China, 2003.

[15] Yang, Y. and Claramunt, C.: A Hybrid Approach for Spatial Web Personalization, Web and Wireless Geographical Information Systems, Lecture Notes in Computer Science 3428, 206-221, Springer-Verlag, 2005.

[16] Nie, Z., Ma, Y., Shi, S., Wen, J. R. and Ma, W. Y.: Web Object Retrieval, Proceedings of the 16th International World Wide Web Conference, 81-90, 2007.

[17] Nie, Z., Zhang, Y., Wen, J. R. and Ma, W. Y.: Object-Level Ranking: Bringing Order to Web Objects, Proceedings of the 14th International World Wide Web Conference, 567-574, 2005.

[18] 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.

[19] Akaike, H.: A New Look at the Statistical Model Identification, IEEE Transaction on Automatic Control, Vol.19, pp. 716-723, 1974.