データベースシステム 128－69
（2002．7．19）

ページ内容と位置情報に基づくWebコンテンツのローカル度検出とその応用

馬　強[††]　松本　知弥子[†]　田中　克己[††]

インターネットやデジタル放送の急激な進歩と普及によって，多くのユーザが多種多様な情報を受信・発信できるようになり，情報資源の量は日々増加し続けている．ユーザが大量の情報の中から，適切な情報を検索することは困難な作業である場合がある．特に，特定のユーザのみが興味を持つ，地域密着情報のようなローカル的な情報を獲得したり，排除するには，従来の情報検索やフィルタリング手法のみでは不十分である場合がある．本論文では，Webページがどの程度地域に密着しているかを計る尺度としてローカル度を定義し，その抽出手法と応用システムについて述べる．また，ローカル度の定義を評価するための予備実験の結果を示す．

Localness Degree of Web Pages and Its Applications from Page Content and Location Information

QIANG MA ,[††] CHIYAKO MATSUMOTO,[†] and KATSUMI TANAKA[††]

The vast amount of information is available on the WWW(World Wide Web). Usually, users use the information filtering technologies or search engines to acquire their favorite information. However, it's still not easy to acquire or exclude local information with the conventional search engines and information filtering technologies. In this paper, we propose a new notion **localness** to discover local information from the WWW. We also propose some useful applications based on localness and show some results of our preliminary evaluation.

1. はじめに

インターネットやデジタル放送の進歩と普及に伴って，ユーザが利用できる情報が大量となっている．特にWWW（World Wide Web）では，ページ数も100億ページに達するようになり，毎日700万ページずつ増え続けている[1),2)]．これらの情報は，専門家のみではなく，多くの一般の人々にも共有・アクセスできるようになっている．

大量の情報から，適切な情報を獲得するためには，検索やフィルタリング手法が有効であり，数多くの研究やサービス[3)~9)]が行われている．従来の検索・フィルタリング手法は，キーワードやユーザプロファイルベースのもの，さらに，リンク構造を用いる手法[10),11)]である．

インターネットを利用するユーザが増えれば増えるほど，ネットワークを日常・地域生活のために利用することが多くなる．つまり，地域情報が活発になるのである．しかし，日常生活・地域に密着する情報をキーワードで記述するするのが困難な場合があるので，従来の検索・フィルタリング手法は，活発になりつつある生活・地域密着情報を獲得したり，排除したりすることが困難である場合がある．

Web上にある地域ポータルサイト[12),13)]には，地域の情報が集められ，地域密着情報はある程度獲得でき，有用であるが，基本的には，人手により収集・登録したものであり，限界がある．これらの地域ポータルサイトの構築・検索や，人手で集めたWebサイトの地域密着度の評価は，Webページの新しい尺度が必要となる．

我々は，局所的な地域名や組織名などの地理情報が多く含まれているページや，場所や時間には依存せずに，いつでもどこにでもある話題は，ローカル度の高い情報と考え，Webコンテンツの地域・生活への密着度合いを測る尺度（ローカル度）とそれに基づく情報フィルタリング機構を提案してきた[14)~17)]．

本論文では，これまでの研究成果をふまえ，Webコンテンツのローカル度の定義とその抽出手法を提案す

† 神戸大学大学院 自然科学研究科 情報知能工学専攻
Department of Computer and Systems Engineering, Graduate School of Science and Technology, Kobe University
†† 京都大学大学院 情報学研究科 社会情報学専攻
Department of Social Informatics, Graduate School of Informatics, Kyoto University

ると共に，実験結果と応用システムを用いてローカル度の有用性と有効性について述べる．

本研究では，次の三つの観点からローカル度の定義と抽出を行う．

- Web ページの内容の地域偏在性
  Web ページ内の地名，組織名などの地理用語の頻度（割合），詳細および内容のカバー範囲の解析を行い，Web ページの内容の地域に固まる（集中）度合いを調べる．つまり，Web ページの内容の地域偏在性に基づいてローカル度の定義と抽出を行う．また，Web ページの内容の地域偏在性を調べる時，そのページのみではなく，その他のページとの比較も行う．
- Web ページの話題の遍在性
  日常な話題は，何時でも何処でもありうる情報である．そのような情報は，特定の地域のユーザにしか興味をもたれない可能性が高く，ローカル度の高い情報である．本研究では，まず，ページ内の日常用語パターンの出現頻度を調べる，この頻度が高ければ，日常な情報であり，ローカル度が高いと考える．同時に，他のページとの比較を行う．場所と時間が異なるが，それ以外の内容が類似しているページが多ければ，それらのページは，日常な話題を取り扱っている可能性が高いと考え，ローカル度の高い情報であるとする．
- ユーザの地域偏在性
  ある地域に固まったユーザしか発信・アクセスしないページは，その地域に密着する情報，つまり，ローカル度の高い情報である可能性が高い．そのため，本研究では，アクセス・発信ユーザの地域分布を調べてローカル度の定義と抽出を行う．

既存のシステムと比較して，本研究で提案するローカル度は，次のような特徴がある．

- 日常生活・地域の密着情報の獲得・排除のための意味的尺度である．本研究で提案するローカル度を利用すれば，ローカルな情報の獲得のみではなく，排除することも可能である．また，キーワードなどによる複雑な質問記述が必要ではないため，誰でも手軽に利用できる．
- 他のページと比較してローカル度を計算しているため，相対的概念である．そのため，既存の地域ポータルサイトでは，獲得困難な複数地域にまたがる情報を獲得可能となる．
- 地域ポータルサイトの構築，評価を行うための客観的な尺度として利用できる．

以下，本論文の構成を示す．まず，2 章で関連研究を示す．3 章では，Web ページのローカル度の定義とその抽出手法について述べる．4 章では，実験結果と考察を述べる．5 章では，ローカル度を用いた応用システムについて述べる．6 章では，まとめと今後の課題について述べる．

2. 関連研究

馬[18),19)]，宮崎[20)] らはニュース記事など Web ページの新規性に着目して，新鮮度・流行度といった概念を提案している．新しい意味的な尺度を用いた情報フィルタリングの試みを行っている点は，本研究と同じである．本研究では，時間のみではなく，空間（地域，位置など）もを考慮している点が異なる．

どの地域の情報であるかという視点からの研究が数多く存在するが，どの程度地域に密着しているかという観点からの研究は，著者らの知る限りでは，未だにない．

モーバイルインフォサーチ[21)] では，ある特定の場所に関する情報をネットワーク上から収集・選択・加工し，ユーザの状況を考慮し，ユーザに適切な情報を提供する「位置指向の情報統合」を目的としている．そして，ネットワーク上のバーチャルワールドと，リアルワールドの双方向の情報を効率良くやりとりし，現在地に関する情報や，世の中の流行，お勧めをユーザに提供する実験[22)] を行っている．ローカル度の抽出手法では，一部位置情報を利用しているが，ローカル的な Web コンテンツを発見するための手法を論じている点が異なる．

Stanford 大の Buyukkokten らの研究[23)] では，Web 上のページが，実際の地理上にどのように分布しているかサーバの IP アドレスや，ページの内容の地理情報，郵便番号などのデータベースから求め，地図上に視覚的にプロットするシステムを作り，Web のページがどの地域に密着しているか分かるようにしている．本研究では，ローカル度を定義するときに，地理用語の割合と位置情報，話題の日常性を調べて，ローカル的な情報を獲得したり，排除するためのローカル度を定義している点が異なる．

デジタルシティ[24)] では，地域コミュニティに向けた情報サービスを実現するため，WWW を地理情報システムによって拡張した拡張 Web 空間とその検索言語を提案している．平松氏らは，拡張 Web 空間とその検索言語をデジタルシティ京都に適用し，WWW に基づいた地理情報サービスを構築している．京都大学の井上氏らの研究[25)] では，Web ページ間のリンク構

造を分析することにより，Webページと地域の関係を評価する手法を，そのページの地域における人気度と，地域志向性から提案している．Geographic Search[26]で，Daniel Egnor氏は，Google[8]を使って，地域を限定した検索ができる方法を実装した．例えば，「家の近所にある全ての書店を検索する」といった用途に使える．このためにページの中に記されている住所などの地理的な情報を，米国が国勢調査のために公開している無料の地理データベースと照らし合わせることによってWebの検索手法を2次元の地図にプロットし，特定の地域のページだけを検索結果として返す方法を実装した．本研究では，リンク構造ではなく，ページの内容から地理用語の割合と位置情報，話題の日常性を調べ，ローカル度を定義して，内容がどれぐらいローカルであるかを応用しようとしている点が異なる．

3. ローカル度

本節では，ページの内容の偏在性，話題の遍在性とユーザの偏在性から様々なローカル度の抽出手法を述べる．これらの抽出手法は別々の利用のみではなく，統合して利用することも可能である．

3.1 内容の地域偏在性によるローカル度

ページ内およびサイト内他のページの地理用語，組織名など固有名詞の割合，詳細と地理範囲に基づいてローカル度を計算する．

(a) ページ内の解析

ローカル度を，ページ内に出現する，ローカルさを特徴付ける地理用語の頻度・詳細度と，その地理用語の位置情報に基づいて定義する．☆

より多く，詳細な地理用語・組織名を含んでいるページは，よりローカル的であると考えられる．

(a-1) 地理用語の頻度・詳細度

一般的に地域密着性の高いWebページは，国・県・市・町村名などの地理情報を表す地理用語を多く含んでいる．すなわち，記事に含まれている地理用語の割合が大きければ，ページのローカル度が高い．小さければ，ローカル度が低いと考えられる．地理用語と同時に，本研究では，組織名をも考慮している．組織名は，会社名・学校名・団体名などである．「市役所」などの組織は，地域に依存して存在するものなので，組織名を多く含んだページもローカルさに関係があると考えるからである．

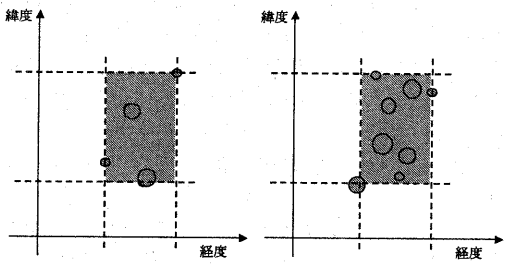

図1 地理用語の詳細度を考慮した密度の例
Fig. 1 Exapmle for Density Based Localness

「姫路」を多く含んでいるページは，「日本」を多く含んだページよりもローカルであると考えられる．従って，地理名詞のレベルを考慮してローカル度を計算することが必要であると思われる．つまり，ページ内に詳しい地名が書いてあるものは，おおまかに書いてあるものよりローカル度が高いと考える．そこで，場所の範囲によってi番目に出てくる地理用語:$geoword_i$には詳細度:weight($geoword_i$)をつける．詳細度は，地域国名 < 組織名 < 地域一般名☆☆の3段階でつけている．地域一般名の中でも県名とその他の詳細度は変えている．地域ごとの人口比なども考慮する必要があると考え，政令指定都市の詳細度は県名と同じにしている．

地理用語の頻度・詳細度によるページpのローカル度$local_g$は次のように計算される．

$$local_g(p) = \frac{\sum_{i=1}^{n} \mathrm{weight}(geoword_i)}{\mathrm{words}(p)} \qquad (1)$$

ただし，$weight$は地理用語$geoword_i$の詳細度であり，$words$はページpの単語の総数（stop wordsを省く）である．nはp内の地理用語の数である．

(a-2) 地理用語カバー範囲と密度

一般的に，局所的な地域の話題の記事の場合，文章に出てくる地理名詞も，局所的な地域のものである．反対に，広い範囲に渡る話題の記事の場合，文章に出てくる地理名詞も，広い範囲に渡って現れることが多い．

そこで，ページ内の地理名詞を，緯度・経度データに従って地図上にプロットした時に，全ての点を含む最小の範囲の面積を調べ，ページの話題のカバー範囲を評価する．この面積が小さいほど，話題のカバー範囲が狭く，ローカル度が高いと考える．

☆ ページ内のトピックが常に1つとは限らず，複数のトピックが存在することもある．本稿では，複数のトピックによりローカルな情報が多く集められて，全体的には広範囲に渡っても，グローバルなページとは見なさず，ローカル度は大きいとする．本論文では，特別な記述がない限り，1ページ1トピックとする．

☆☆ 郵便住所の町名までのデータ

この時，最小外接矩形:MBR（Minimum Bounding Rectangle）[27),28)] を利用する．本研究では，x 軸を緯線の方向，y 軸を経線の方向として MBR を作成する．MBR の面積が大きければ，広い範囲にわたる記事であることを示し，ローカルさは小さい．反対に小さければ，局所的な地域の話題なので，ローカルさは大きい．

しかしながら，**図 1** で示されているように，面積が同じでも，プロットされる地理用語の数，詳細などが異なると，ページのローカルさは異なる可能性が高いと考える．図 1 の左では，MBR の内側に 4 個の点が含まれており，右では，8 個の点が含まれている．左の例より右の例の方が詳細度の高い点の数が多く，地域に密着している可能性が高いので，ローカル度が高いと考える．つまり，MBR に含まれる点が多く，詳細度に基づいた点のサイズが大きい場合，局所的な地域密着情報であるため，ローカル度が高いと考える．従って，MBR の面積だけでは，不十分であり，MBR の中に何個の点がプロットされているかや，その点が示す地理用語の詳細度も考慮することが，ページのローカル度を定義するには必要であると思われる．

本研究では，ページ内に出現する地理用語を地図上にプロットしたときに，それらを全て内側に含む最小の長方形:MBR を使って求めた面積の中に，どれぐらいの詳細度の地理名詞がプロットされているかという，密度を求める．面積が小さく，その面積内に詳細度の高い地理名詞が集中しているほど，局所的な話題である可能性が高いので，密度が大きいほど，ローカル度が高く，密度が小さいほど，ローカル度が低いと考える．

ページ p の地理用語 $geoword_i, 1 \leq i \leq n$ のカバー範囲 MBR での密度によるローカル度 $local_{de}$ は次のように計算される．

$$local_{de}(p) = \frac{\sum_{i=1}^{n} \mathrm{weight}(geoword_i)}{MBR(p)} \qquad (2)$$

ただし，n はページ p 内の地理用語の数である．

緯度・経度のデータとしては，全国 3252 個の都道府県庁，市区役所，町村役場の所在地データを用いる[29)]．それよりも細かい地名の場合は，1 つ上のレベルに上げて，その場所の地名を用いる．例えば，「兵庫県姫路市大津区 …」の場合，兵庫県姫路市の市役所所在地のデータを用いる．

記事の中の地理名詞を緯度・経度データと照合する場合，全国には同じ名前の地名がたくさんあるという問題がある．しかし，ほとんどの場合において最初に地名が出てくる時は，○○県●●市や，☆☆県★★町という風になっているので，絞り込むことができる．また，既出の地理名詞の場合には突然，文書の中に◇◇市と出てくる場合があるので，この場合は，既出の地理名詞かどうかを調べて判断する．

(b) 相対的ローカル度：比較

同一サイト内の他のページには詳細な地理用語が少なければ，多くの詳細な地理用語を含むページのローカル度が高いと考える．サイト内のすべてのページが多数の詳細な地理用語を含んでいれば，そのサイトでは地理用語はあまり重要ではない可能性が高く，地理用語の頻度・詳細度などを用いて計算されるページの特徴量：ローカル度を下げるべきであると考える．すなわち，ローカル度は相対的である．

ある Web サイト内，$local_i$[⋆]が閾値 θ より大きいページの数は m とする．ページ p の相対的ローカル度 $local_c$ は，次のように計算される．

$$local_c(p) = local_i(p)/m \qquad (3)$$

3.2 ページ内の話題の遍在性によるローカル度

(a) 日常性の高い話題

日常生活情報など日常性の高い話題は，何時でも何処でもありうる情報なので，特定の地域の人々にしか興味を持たれない可能性が高く，ローカル度が高いと考える．たとえば，日常性の高いスーパーの特売情報は，そのスーパーの近所の人々しか興味を持たないので，ローカル度は高いと思われる．

夏祭りなど，場所に依存せず普遍的にどこにでもある同様のイベントの情報は，日常的なものであり，そういった情報を発信している Web ページは，最も興味を持たれるのはその地域であると考えられるので，ローカル度が高い．つまり，時間・場所と関係なく，どこでもありうるイベントに関する情報は，日常的話題であり，特定の地域のユーザしか興味を持たない可能性が高いと考え，ローカル度の高い情報とする．

(a-1) ページ内の日常用語パターンの頻度

ページ内の日常用語（パターン）⋆⋆の割合が高ければ，ページの日常話題である可能性が高く，ローカル度が高い．

本研究では，日常生活情報のページ群から，共起度の高い単語のパターンを抽出し，日常用語パターンの辞書を生成する．例えば，「スーパー」と「米」が同時に出現する場合は，「スーバー」と「米」は日常用

⋆ $local_{de}$ と $local_g$ を統合したものである．つまり，$local_i(p) = f(local_{de}(p), local_g(p))$ である．例えば，$local_i(p) = \alpha \cdot local_{de}(p) + (1-\alpha) \cdot local_g(p)$．$\alpha \in [0,1]$ は重みである．

⋆⋆ 本論文では，日常生活情報を述べる文書でよく出てくる固有名詞以外の単語を日常用語とする．

語として使われている可能性が高いと考えられる．一方，「米」，「WTO」と同時に出現していれば，「米」と「WTO」は日常用語として使われている可能性が低いと考える．この辞書を用いてページ p の日常用語パターン☆の出現頻度を調べる．出現頻度が高ければ，ページ p は日常性が高く，特定の地域のユーザにしか興味持たない可能性が高い．つまり，ローカル度が高い．

$$local_{fpd}(p) = \frac{2 \cdot \sum_{i=1}^{n} wp_i}{words} \quad (4)$$

ただし，$local_{fpd}$ はページ p 内の日常用語パターンの頻度によるローカル度を表す．wp_i は p 内の日常用語パターン i の出現回数であり，n は日常用語パターンの総数である．$words$ は，p の単語総数である．

(a-2) 他のページとの比較による日常性

日常性の高い情報（Web ページ）は，その他の情報（ページ）と比較すると，地名や時間など固有名詞が違いが，それ以外の内容で類似している可能性が高い．そこで，本論文では，Web ページの間における，地理，時間など固有名詞の比較と，それらを省いた内容部分の比較を行う．固有名詞が異なるが，内容部分の類似するページを多数有するページは，日常情報を発信している可能性が高いので，ローカル度の高い情報であると考える．

時間と場所が異なるが，ページ p の内容部分と類似するページの数が m とする．ページ p の他のページとの比較による日常性に基づくローカル度 $local_{mud}(p)$ は，次のように計算される．

$$local_{mud}(p) = m/n \quad (5)$$

ただし，n は，比較ページの総数である．

(b) 注目度の高い話題

あるイベントに対して，異なる多くの Web サイトから報道配信されていれば，そのイベントはグローバルである可能性が高く，そのイベントを発信しているページ（群）は，注目度の高い，ホットな話題であり，ローカル度が低いと考える．

他のページとの比較による日常性に基づくローカル度 $local_{mud}$ では，時間・場所の不一致であるが，内容部分が類似する他のページを求めている．これに対して，ここでは，内容類似のみではなく，時間・場所も一致している類似ページを求めている．このような類似ページを多数有するページは，注目度の高いイベントを発信している可能性が高いので，ローカル度が低いと考える．

☆ $(word_1, word_2)$ のような単語のペアで定義される

$$local_{hot}(p) = 1 - m/n \quad (6)$$

ただし，$local_{hot}$ はページ p の話題の注目度によるローカル度である．n は比較ページの総数である．m は，p と，時間・場所が一致かつ内容が類似しているページの数である．

3.3 ユーザの地域遍在性によるローカル度

(a) 利用者の地域偏在性

ページをアクセスしているユーザは，ある地域に固まっていれば，そのページはその固まった地域にいるユーザしか興味持たさない可能性が高いので，ローカル度が高いと考える．そのため，本研究では，ページ p のアクセスログを用いてユーザの地域分布を調べてローカル度の計算を行う．

p をアクセスしたユーザの IP を地図上にプロットして，クラスタリングを行い，それぞれのクラスタのカバーする範囲の面積を求める．もし，多くの IP は，一個のカバー範囲（面積）が小さいクラスタにプロットされていれば，ページ p のローカル度が高いとする．つまり，多くの点（IP）がある狭い範囲にプロットされていれば，そのページはある特定場所のユーザにしか興味持たないので，ローカル度の高い情報であると考える．

そこで，IP クラスタ c_i のカバー範囲の面積 $area(c_i)$ と，c_i にプロットされている IP の割合に基づいてページ p のローカル度 $local_{ip}$ を次のように計算する．

$$local_{ip} = \max((c_i)/area(c_i)) \quad (7)$$

ただし，$1 \leq i \leq n$, n はクラスタの数である．

(b) 発信者の地域偏在性

あるイベントを発信しているすべてのユーザがある地域に固まっていれば，その情報は地域のイベントに関するものである可能性が高く，ローカル度が高いと考える．

そのため，本研究では，ページ p と，時間・場所が一致かつ内容が類似であるページを発信するユーザの IP を収集し，地図上にプロットして，発信者の地域分布を調べてローカル度の計算を行う．そのとき，利用者 IP の地域分布と同様の手法を用いる．ローカル度の計算も式 (7) を利用する．

4. 実験と考察

ローカル度の定義を評価する予備実験を行った．

4.1 実　　験

実験には，ASAHI.COM（ `http://www.asahi.com/` ）の記事を使用した．各記事の HTML 文書ソースには，記事の始まる前後に，<!- - Start of kiji - ->，<!- - End of kiji - -> の記述があるので，その内側に

表 1　ローカル度の定義方法
Table 1　definitions of localness degree

	地理用語	日常語
ページ内処理	地理用語の密度:$local_{de}(p)$ - 地理用語の頻度・詳細度・位置情報	日常語の共起度パターン:$local_{fpd}(p)$ - 共起しているパターンから日常語辞書作成
他のページとの比較処理	相対的ローカル度:$local_{c}(p)$ - 地理用語の重要さを相対的に判定	比較による日常性:$local_{mud}(p)$ - 類似しているページ数から日常性を判定

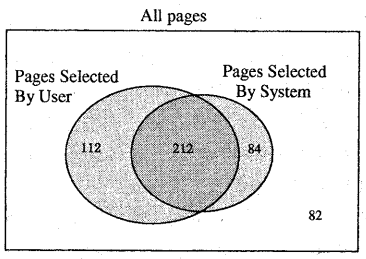

図 2　定義式 (2) による実験結果
Fig. 2　Results of Preliminary Evaluation: Case of Function (2)

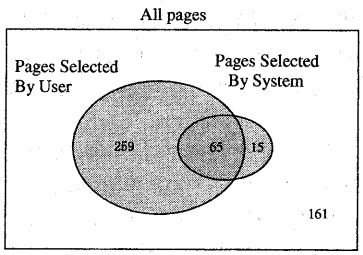

図 3　定義式 (5) による実験結果
Fig. 3　Results of Preliminary Evaluation: Case of Function (5)

書かれた記事だけを形態素解析の対象にし，それ以外のサイト内へのリンクや，広告の記述は外した．

予備実験 1 では，`http://www.asahi.com/` の記事の約 500 ページのローカル度をページ内の地理用語の頻度から，式 (2) により計算し，筆者の判断で正解記事を選び，適合率と再現率を求めた．システムの正解は，ローカル度 10 以上とした．その結果，再現率は 0.654 で，適合率は 0.716 という結果になった．

予備実験 2 では，同じ実験材料で，ローカル度を式 (5) を用いて，場所に関係なく普遍的に存在する日常的な話題を考慮して求めた．その結果，再現率は 0.200 で，適合率は 0.813 という結果になった．

図 2，**図 3** に，実験の結果を示す．実験材料全記事のうち，筆者が判断したローカルな記事が左の円，システムがローカルだと判断した記事が右の円である．それぞれの内訳の記事数を示している☆．

4.2 考　　察

表 1 に，本論文で述べてきたローカル度の定義についてまとめた．地理用語と，日常用語それぞれについて，ページ内処理をしたものと，比較による処理をしたものの二つがある．

この内，Lcl_{de} と Lcl_{mud} について実験を行った．予備実験 1 では，ある程度の精度が得られ，ローカル度の判定に用いることができると考えられる．予備実験 2 では，再現率が低くなっている．これは，実験に用いたデータが約 1 週間分という短い期間なので，筆者がローカルであると判断した内容と，類似した内容が比較対照の中に存在しなかったからだと考える．これは，長い期間のデータを集めて類似度を計算すると，解決することができると考える．

また，これらの定義は，ユーザの目的によって組み合わせて使い分けることができると考える．

5. 応用システム

本節では，ローカル度を用いた応用システムについて述べる．

5.1 ローカル度を用いたフィルタリング機構

検索エンジンなどで，検索窓にキーワードを入力しただけでも，ローカルな情報を得ることは可能であるが，キーワードの指定が困難であったり，どれぐらいローカルな情報が欲しいか指定することは難しい．つまり，キーワードのみで，ローカルな情報を獲得するための問い合わせ記述をするのが困難である場合がある．また，ローカルな情報を獲得するだけでなく，排除したいという要求には応えることができない．

そこで，ローカル度の値によるフィルタリングを行う．ローカルな情報を得たい場合は，ローカル度の高いページを検索結果の上位にランキングするようにする．反対に，ローカルに依存した情報を排除したい場合は，ローカル度の高いページを検索結果から外す．

☆ 人間の判断基準は，特定地域の人にとってのみ有用と考えられる情報をローカルとしている．システムは，それぞれの定義式によって計算された値が閾値より大きいページかどうかによって判断している

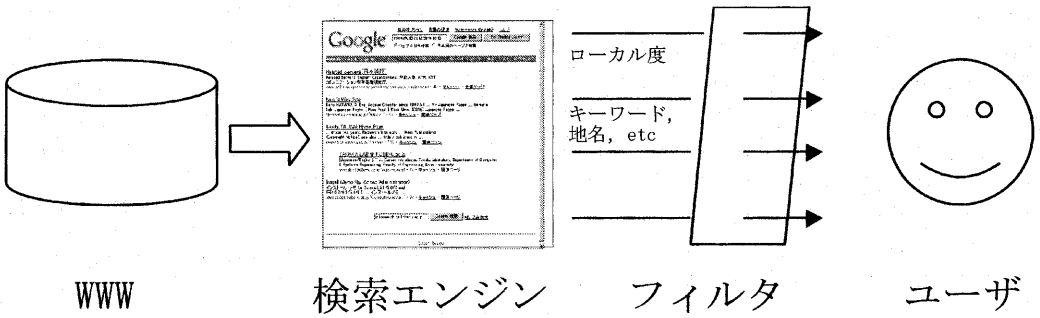

図 4 フィルタリングのモデル図
Fig. 4 Filtering Model Based on Localness

フィルタリングモデルを図 4 に示す．ユーザが検索質問をすると，WWW サーバから検索エンジンが情報を検索し，その結果からローカル度を用いてフィルタリングを行い，ローカルな情報を獲得したり，排除したりする．

このフィルタリング機構によって，従来の検索エンジンやフィルタリング手法では，困難であった，

- 特定の地域・組織に密着したページを除いたページ集合を求める
- 検索する対象地域を限定せずに，知りたい情報が依存している地域密着型のページ集合を求める
- 複数地域の地域密着情報が集まったページから情報を検索する

などが可能になると思われる．

5.2 モバイル環境におけるローカル度を用いた地域放送コンテンツの動的再構成

i-mode を代表とする携帯通信端末の普及に伴い，モバイル環境での情報共有，アクセスは活発化しつつある．また，次世代の携帯通信端末では，映像コンテンツなどもを快適にアクセス可能であると予測される．つまり，放送コンテンツや Web コンテンツなどを携帯通信端末で容易にアクセス可能となる．一方，ブロードバンドや無線通信技術の発達と普及に伴い，地域放送などによる地域情報の発信もより活発化してきている．

そこで，モバイル環境における地域放送コンテンツの効率アクセスのため，我々は，ユーザの位置情報 (Location Information) の変化とローカル度の相対性を利用して，複数の地域の放送コンテンツを動的に再構成して，ユーザに呈示する機構を提案する．アイディアは次のようである：

- 地域放送の基地局とユーザの距離に応じて，それぞれの地域放送から受信するコンテンツの割合を動的に決める．たとえば，大阪，神戸と京都の三つの地域放送局とユーザの距離比は 1:2:3 とすると，三つの局から受信するコンテンツの比も 1:2:3 とする．
- ユーザがあらかじめ定義した受信コンテンツのローカル度のレベルに応じて，受信コンテンツの動的構成を行う．たとえば，大阪のローカルコンテンツ 4 割，近畿のローカルコンテンツを 3 割，全国コンテンツを 3 割という指定することが可能である．ユーザの位置情報に応じた受信コンテンツの配分は動的であるが，ローカルレベルに応じた受信コンテンツの配分は静的である．

6. おわりに

本論文では，地域や日常生活に特化した，ローカルな情報を獲得したり排除したりすることを可能にするためにローカル度という Web ページの特徴量を定義し，それを用いたフィルタリング機構などの応用システムを提案している．

本論文では，特定の地域・組織に依存したユーザのみが興味・関心を持つ，日常生活・地域の密着情報はローカルであると考え，次の三つの観点からローカル度の定義と抽出を行っている．

- ページ内容の地域偏在性
- ページの話題の地域遍在性
- 発信・アクセスユーザの地域偏在性

予備実験で，`http://www.asahi.com/` の記事のローカル度を地理用語の密度を考慮して計算すると，再現率は 0.654 で，適合率は 0.716 という結果になった．ある程度の精度は得られ，情報検索，フィルタリングの基準とすることができると考えられる．

今後は，ローカル度の抽出アルゴリズムの改良や実験検証などを行う予定である．また，提案した応用システムのプロトタイプシステムを作成して，提案する概念と応用システムの有効性と有用性を検証する予定

である．

謝辞 本研究の一部は，平成 14 年度文部科学省科学研究費特定領域研究 (2)「Web の意味構造に基づく新しい Web 検索サービス方式に関する研究」(課題番号：14019048，代表：田中克己)，および特定領域研究 (A)(2)「モバイル環境におけるコンテンツのマルチモーダル検索・呈示と放送コンテンツ生成」(課題番号：14208036，代表：田中克己) によっております．ここに記して謝意を表すものとします．

参 考 文 献

1) Cyveillance: http://www.cyveillance.com/.
2) InternetLibrary: http://www.archive.org/.
3) Yahoo!Japan: http://www.yahoo.co.jp/.
4) goo: http://www.goo.ne.jp/.
5) LycosJapan: http://www.lycos.co.jp/.
6) Infoseek: http://www.infoseek.co.jp/.
7) Excite: http://www.excite.co.jp/.
8) Google: http://www.google.com/.
9) WiseNut: http://www.wisenut.com/.
10) Kleinberg, J. M.: Authoritative sources in a hyperlinked environment, *Journal of the ACM*, Vol. 46, No. 5, pp. 604–632 (1999).
11) Chakurabarti, S., Dom, B., Gibson, D., Kleinberg, J. M., Kumar, S. R., Raghavan, P., Rajagopalan, S. and Tomkins, A.: Hypersearching the Web. Scientific American (1999).
12) Yahoo!地域情報: http://local.yahoo.co.jp.
13) まち goo: http://machi.goo.ne.jp.
14) 松本知弥子, 馬強, 田中克己: 地理情報を用いたニュースのフィルタリング機構, 情報処理学会第 62 回全国大会, Vol. 3, pp. 505–506 (2001).
15) 松本知弥子, 馬強, 田中克己: Web ページのローカル度検出に基づく情報フィルタリング, 情報処理学会研究報告 (DBWS2001), Vol. 125, No. 36, pp. 273–280 (2001).
16) 松本知弥子, 馬強, 田中克己: Web ページの地理情報と話題の日常性を考慮したローカル度検出とフィルタリング機構, データベースと Web 情報システムに関するシンポジウム (DBWeb2001), IPSJ Symposium Series, No. 17, pp. 193–200 (2001).
17) Matsumoto, C., Ma, Q. and Tanaka, K.: Web Information Retrieval Based on the Localness Degree, *DEXA*, Lecture Notes in Computer Science, Springer (2002).
18) 馬強, 角谷和俊, 田中克己: 放送型情報配信システムのための時系列性を考慮した情報フィルタリング, 情報処理学会論文誌:データベース, Vol. 41, No. SIG6(TOD7), pp. 46–57 (2000).
19) Ma, Q., Miyazaki, S. and Tanaka, K.: WebSCAN: Discovering and Notifying Important Changes of Web Sites, *DEXA*, Lecture Notes in Computer Science, Vol. 2113, Springer, pp. 587–598 (2001).
20) 宮崎慎也, 馬強, 田中克己: WebSCAN:Web サイトの変更発見と放送型変更通知, 情報処理学会論文誌:データベース, Vol. 42, No. SIG8(TOD10), pp. 96–107 (2001).
21) 三浦信幸, 高橋克巳, 横路誠司, 島健一: 位置志向の情報統合～モーバイルインフォサーチ 2 実験～, 情報処理学会第 57 回全国大会, Vol. 3, pp. 637–638 (1998).
22) MIS2 (モーバイルインフォサーチ 2 実験): http://www.kokono.net/.
23) Buyukkokten, O., Cho, J., Garcia-Molina, H., Gravano, L. and Shivakumar, N.: Exploiting Geographical Location Information of Web Pages, *WebDB (Informal Proceedings)*, pp. 91–96 (1999).
24) 平松薫, 石田亨: 地域情報サービスのための拡張 Web 空間, 情報処理学会論文誌:データベース, Vol. 41, No. SIG6(TOD7), pp. 81–90 (2000).
25) 井上陽介, 李龍, 高倉弘喜, 上林弥彦: 地域ウェブ情報検索のためのリンク構造分析によるウェブページと地域の関係抽出, *Data Engineering Workshop* (2002).
26) Egnor, D.: Geographic Search (2002). `http://www.google.com/programming-contest/winner.html`.
27) Guttman, A.: R-trees: A dynamic index structure for spatial searching, *Proc. ACM SIGMOD Conference on Management of Data*, Vol. 14, No. 2, pp. 47–57 (1984).
28) Zaniolo, C., Ceri, S., Faloutsos, C., Snodgrass, R. T., Subrahmanian, V. S. and Zicari, R.: *Advanced Database Systems*, The Morgan Kaufmann (1997).
29) 武田尚志: 全国都道府県市町村・緯度経度位置データベース 1.0 (2000).