地域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

Computation of Relationships between Web Pages and Locations by analyzing Link Structure

井上陽介[†] 李龍[††]
高倉弘喜[†††] 上林弥彦[††]

Yohsuke Inoue,[†] Ryong Lee,[††] Hiroki Takakura[†††]
and Yahiko Kambayashi[†]

本論文で、ウェブページ間のリンク構造を分析することにより、ウェブページと地域の関係を評価する手法を提案する。既存の検索システムでは、地域と無関係に単純に人気のあるページが上位にランクされるため、地域情報検索ではウェブページがどの地域とどのようなの関連性をもっているか評価することが必要となる。また、ウェブページのリンクをリンク先のサイトを推薦する制作者の意志と評価することができる。そのため、どのような地名を含むページからリンクされているか調べることにより、そのページの地域における人気度を知ることができる。また、リンク先のページがどのような地名を含むか調べることにより、そのページの地域指向性を知ることができる。我々はウェブページのリンク構造からウェブページの地域からの人気度と地域への指向性を求めることの有用性を示し、さらにその際に起こる問題点にふれ効果的な計算手法についても提案する。

1. はじめに

インターネットの普及により、世界中の情報を誰もが容易に手に入れることができるようになった。しかし、特定の地域に関連した情報に対する需要は依然として大きい。また、その一方で情報が膨大になったために、そこから地域情報だけを選択して収集することが困難となった。

そこで、我々は “ウェブ上の情報” を “地理情報システム (Geographical Information System , GIS)” と統合して利用することにより、効率的に特定の地域に関する情報検索を支援するシステム KyotoSEARCH[1)] を開発している。

一般に地名は文章の中で省略されることが多い。そのため、地域情報を扱ったウェブページであっても、その地域の地名が繰り返し出現することは少なく、従来のようなキーワード中心の手法だけでは、そのウェブページが地域情報を持つかどうかを判断するのは困難である。そこで我々は地域情報検索の手段として、ウェブページが含むキーワードの分析に加えて、ウェブページ間のリンク構造を利用することに着目した。

既にリンク構造を利用したウェブ上の情報検索手段が実用化されている。これらのシステムの多くは、リンク構造から計算した人気度でウェブページをランク付けする。ウェブページにリンクを張ることをリンク先のウェブページを推薦する行為と見なすことにより、多くの推薦を集めるウェブページは良質なウェブページであるという仮定に基づいている。この手法はすでにウェブ情報検索システムにおいて高い評価を得ている。

しかし、この人気度は世界中のウェブページからのリンク構造にのみ依存し、ウェブページに含まれる地域情報・地名キーワードをまったく考慮していない。そのため、人気度が高いウェブページが必ずしもその地域に関連した情報を持つとは限らないという問題点を持つ。

そのため、例えば次のようなケースにおいて問題が

† 京都大学工学部
Faculty of Engineering, Kyoto University
yohsuke@db.soc.i.kyoto-u.ac.jp
†† 京都大学情報学研究科社会情報学専攻
Department of Social Informatics Graduate School of Informatics, Kyoto University
{ryong , yahiko}@db.soc.i.kyoto-u.ac.jp
††† 京都大学 大型計算機センター 研究開発部
Data Processing Center, Kyoto University
takakura@rd.kudpc.kyoto-u.ac.jp

生じている。New York Times は、全米のサイトからリンクされているため高く評価され、結果として “New York” というクエリに対して上位にランク付けされる。しかし、このウェブページはアメリカの代表的な全国紙のものであり、New York の地域情報としては必ずしも適切な情報ではない [4]。

そこで地域情報検索を効率的に行うために、ウェブページがどの地域と関連しているかを考慮した新しいランキング手法が必要となる。

本論文では、そのために必要なウェブページと地域との間の関係をウェブページ間のリンク構造から評価する手法を提案する。

提案する手法では、ウェブページと地域の間の関係は、ウェブページがどれだけその地域から評価されているか（**ウェブページの地域における人気度**）、またどれだけその地域を意識した内容を持っているか（**ウェブページの地域に対する指向性**）、という 2 つに分類して考える。

本論文では、ウェブページの内容解析とウェブページ間のリンク構造の分析により、ウェブページの地域における人気度とウェブページの地域に対する指向性を評価できることを示し、両者を統合して利用することで効率的に地域情報検索を行う手法について考察する。

以下、本論文の構成を述べる。まず、2 章では、ウェブページの地域における人気度の計算手法を提案し、既存の手法との違いや計算の際に起こる問題点と効果的な計算手法について述べる。3 章ではウェブページの地域に対する指向性を評価する方法について述べる。4 章では 2 章、3 章で提案した方法で実際に計算し、実験方法および結果について説明し、考察する。6 章では関連研究の紹介を行う。7 章で、まとめと今後の課題について述べる。

2. ウェブページの地域での人気度

2.1 ウェブページの人気度計算

L. Page 氏らの研究 [5] で、ウェブページ間のリンク構造から各ページの人気度を計算する方法がある。この方法は既に実用化されて、高い評価を得ている。基本的な考え方は、リンクはリンクされたウェブページへの推薦ととらえることにある。すなわち、多くのウェブページからリンクされているウェブページほど品質がよいと考えられる。さらに、重要なウェブページからリンクされているウェブページはやはり重要であるとしており、この点で単なる被リンク数による評価とは異なる。

この人気度を計算するためにユーザの行動を次のようにモデル化する。

- 各利用者は 1 つのページに単位時間とどまり、ある確率でリンクをたどって次のページに移る

このモデルの元で、多くのユーザがウェブを利用しているという前提で、十分な時間が経過した後により多くの人が見ているウェブページほど品質のよい人気のあるウェブページであると考えることができる。この方法は、次のような評価を同時に下している。

- 人気度の高いページからのリンクを高く評価する
- 総リンク数の多いページからのリンクを軽視している
- 人気度の低いページ同士の相互リンクだけでは評価があがらない

このようなモデルを考えることにより、人気度計算は 1 重マルコフ連鎖の定常確率を求める計算に帰着する。

すると、ウェブページ w_i の人気度を $a(i)$、n 回の遷移後のウェブページ w_i の人気度を $a_n(i)$ とすると、$a(i)$、$a_n(i)$ は次のように表現することができる。ただし、$p(w_i \to w_j)$ はウェブページ w_i からウェブページ w_j への遷移確率を表す。

$$p(w_j \to w_i) = \begin{cases} \dfrac{w_j \to w_i \text{ のリンク数}}{w_j \text{ からのリンク数}} \\ 0 \quad (w_j \text{ からのリンクがない場合}) \end{cases}$$

$$a_{n+1}(w_i) = \sum_{w_j} p(w_j \to w_i) a_n(w_j)$$

$$a(w_i) = \lim_{n \to \infty} a_n(w_i) \qquad (1)$$

ただし、この計算はウェブページのリンク構造が強連結になっていることが前提となっていることに留意しなければならない。上記の数式を強連結集合に適用した場合、各ウェブページの人気度は 0 でない一定の値に収束し、その和は一定となることが知られている。

しかし、現実のウェブ空間から得られるリンク構造は強連結ではない。そこで、Page 氏らの方法ではこの問題に対処するために、ユーザの行動モデルを次のように修正している。

- 現在のウェブページにリンクが存在する場合、各利用者は 1 つのページに単位時間とどまり、確率 α でリンクをたどって次のページに移り、確率 $1-\alpha$ でリンクとは関係なくランダムに他のウェブページに移動する
- 現在のウェブページにリンクが存在しない場合、各利用者は 1 つのページに単位時間とどまり、ランダムに他のウェブページに移動する

最後の条件は、“リンクを 1 つも持たないウェブページ” を、“すべてのウェブページへのリンクを持つウェ

ブページ” と読み替えることに相当する。このようなモデルを採用することにより、前述の計算式 (1) を修正した人気度計算の式は次のようになる。

$$p(w_j \to w_i) = \begin{cases} \dfrac{w_j \to w_i \text{ のリンク数}}{w_j \text{ からのリンク数}} \\ \dfrac{1}{\text{全ページ数}} \quad \left(\begin{smallmatrix} w_j \text{ からの} \\ \text{リンクがない場合} \end{smallmatrix}\right) \end{cases}$$

$$a_{n+1}(w_i) = \alpha \sum_{w_j} p(w_j \to w_i) a_n(w_j) + (1-\alpha)$$

$$a(w_i) = \lim_{n\to\infty} a_n(w_i) \qquad (2)$$

2.2 対象を限定したウェブページの人気度計算

2.1 で説明した既存の人気度計算アルゴリズムは多くの場合において大変効果的であり、高い評価を得ている。しかしながら、地域情報検索においてこのアルゴリズムをそのまま適用した場合いくつかの問題が発生する。

既存の人気度計算アルゴリズムでは地域とは無関係に世界中から集めたウェブページ集合全体に対して行っている。しかし、ウェブページの人気度は地域や目的などによって変化するのが一般的である。そのため、地域と無関係に人気度の高いウェブページがたまたまキーワードとして地名を含んでいた場合などに問題が生じる。そこで、人気度計算の対象を地域によって限定することでウェブページの人気度を計算する必要がある。

地域における人気度を定義するために、2.1 で述べたモデルを次のように修正する。

- 多くの人がウェブを利用している
- 現在のウェブページから**対象地域と関連のある**別のページへのリンクが存在すれば、各利用者は現在のページに単位時間とどまった後、確率 α でリンクをたどって対象地域と関連のあるウェブページの 1 つに移り、確率 $1-\alpha$ でリンクと無関係に対象地域と関連のある任意のページにうつる。
- 現在のウェブページから対象地域と関連のある別のページへのリンクが存在しなければ、リンクと無関係に対象地域と関連のある任意のページにうつる。

これは地域によって節点・枝を限定することで得た部分グラフに対して、2.1 の一般的な人気度計算手法を適用することに相当する。

すなわち、地域のウェブページ集合を L とすると、2.1 の各計算式とその値は以下の様に修正される。

$$p(w_j \to w_i) = \begin{cases} \dfrac{w_j \to w_i \in L \text{ のリンク数}}{w_j \text{ から } L \text{ 内へのリンク数}} \\ \dfrac{1}{L \text{ のページ数}} \quad \left(\begin{smallmatrix} w_j \text{ から } L \text{ への} \\ \text{リンクがない場合} \end{smallmatrix}\right) \end{cases}$$

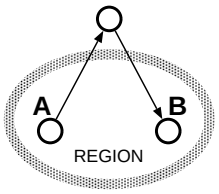

図 1 地域限定で消失する関係

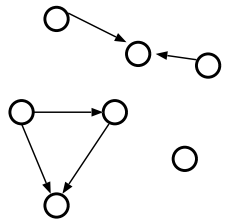

図 2 不十分なリンク構造

$$a_{n+1}(w_i) = \alpha \sum_{w_j \in L} p(w_j \to w_i) a_n(w_j) + (1-\alpha)$$

$$a(w_i) = \lim_{n\to\infty} a_n(w_i) \qquad (3)$$

2.3 部分グラフの拡張

2.2 で述べた方法は単に、地域を限定することで得られる元のリンク構造グラフの部分グラフに対して既存の人気度計算手法を適用するものである。しかし、次のような理由により対象を限定するだけでは十分な精度が得られないと考えられる。

まず、図 1 で示すようなウェブページのリンク構造と地域の限定を考えたとき、図のページ A とページ B の間のような関係が消失して扱われてしまうという点である。しかし、実際にはこのようなリンク関係を持つウェブページは多数存在すると考えられる。従って何らかの形で両者の間の関係をモデルに反映させなければならない。

次に、図 2 で示すように、ウェブページの間にリンク構造が十分に存在しない場合が考えられる。2.1 で論じたような人気度計算モデルはウェブページ間に十分なリンク構造が存在することが前提であり、強連結グラフであることが推奨される。世界中のすべてのウェブページを対象にして人気度計算を行う限りは、強連結グラフに近い状況が得られるが、地域限定によって得られた部分グラフであればリンク構造が不十分な場合も考えなければならない。

そこで、地域を限定した部分グラフを何らかの形で拡張するし、拡張した部分グラフに応じたアルゴリズムの修正が必要となる。本節では次のような部分グラフ拡張方法を提案する。

2.3.1 地域による部分グラフの拡張

前述のような不十分なリンク構造は、多くの場合において行き過ぎた対象限定により、部分グラフが必要以上に小さくなってしまうことに起因する。そこで、対象地域の関連ウェブページだけでなく、対象地域の周辺地域の関連ウェブページを同時に利用することにより、計算に利用できるウェブページが増え、この問題を解決することができる。

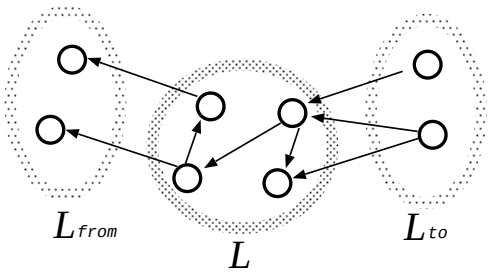

図 3 リンクによる部分グラフの拡張

2.3.2 リンクによる部分グラフの拡張

地域のウェブページとリンクでつながったウェブページは、キーワードを含んでいなくてもその地域とつながりがあると考えられる。そこで地域のウェブページ集合 L に対し、L にリンクしているウェブページ集合 L_{to}、および L からリンクされているウェブページ集合 L_{from} を部分グラフに加えることを考える。

この L_{to}、L_{from} を利用して拡張することにより、単純な部分グラフでの人気度計算の問題点が解消されることが期待される。

3. ウェブページの地域指向性

地域における人気度の高いウェブページが必ずしも地域に密着した内容を持っているとは限らない。そこで、ウェブページがどれだけその地域を意識して作られたか評価するために、“地域指向性” という考え方を導入する。

会話や文章において地名は省略される傾向にあるため、ウェブページ内に同じ地名が繰り返し出現することはまれである。そのため、ウェブページが含む地名を見るだけではそのウェブページが地域情報を提供しているページであるかどうか判断することは難しい。そこで、我々はウェブページが含む地名だけでなく、そのリンク構造を評価することでウェブページの地域指向性を評価することを考える。

あるウェブページがある地域を深く意識して作られている場合、そのウェブページからリンクされているウェブページもやはりその地域を深く意識していると考えられる。したがって、次のようなモデルを考えることによって、そのウェブページの地域指向性を評価することができる。

- あるページを見ているユーザがいる。
- そのページに地域の地名が含まれていれば、ユーザは単位時間だけそのページにとどまり、ある確率でリンクをたどって次のページに移る。
- そのページに地域の地名が含まれていなければ、ユーザはウェブブラウジングを終了する。
- リンクのないウェブページに達した場合も、ユーザはウェブブラウジングを終了する。

このようにした場合に、ユーザがウェブブラウジングを終えるまでに、より多くのページを訪問したページが地域指向性の高いウェブページとみなす。すなわち、平均訪問ページ数で判断することができる。

しかし、上記の条件だけでは、サイト内強連結成分によってユーザの訪問ページ数が発散してしまう場合がある。そこで、前述のユーザのウェブブラウジングモデルに次の条件を付け加える。

- 現在見ているウェブページによらず、ユーザは確率 α $(0 < \alpha < 1)$ でウェブブラウジングを終了する。

また、現実のウェブブラウジングにおいてユーザが永遠にリンクを辿り続けることはないので、この修正はウェブブラウジングモデルとしても妥当といえる。

以上より、対象地域のウェブページ集合を L としたとき、ウェブページ w_i の地域指向性 $b(w_i)$ は次のようにして計算することができる。

$$p(w_j \to w_i) = \begin{cases} \dfrac{w_j \to w_i \in L \text{ のリンク数}}{w_j \text{ からのリンク数}} \\ 0 \quad \left(\begin{array}{c} w_j \text{ からのリンクがない場合} \\ \text{または } w_i \notin L \text{ の場合} \end{array}\right) \end{cases}$$

$$b_0 = 1$$

$$b_{n+1}(w_i) = 1 + \alpha \sum_{w_j \in L} p(w_i \to w_j) b_n(w_j)$$

$$b(w_i) = \lim_{n \to \infty} b_n(w_i) \qquad (4)$$

$b(w_i)$ は単調非減少かつ $1 < b(w_i) < (1-\alpha)^{-1}$ なので、適当な値に収束する。

4. 実験と考察

“京都市” に関連するウェブページを収集し、我々の提案するウェブページと地域の関係抽出手法を評価する実験を行った。

4.1 実験データ

実験に利用するウェブページ集合は、Pentium III 800MHz 1GB メモリを備えた 3 台のマシンからなるシステムを用いて、「京都市」に関連するキーワードを含むウェブページを約 120,000 ページ集めて生成した。収集したウェブページからはリンク構造を抽出し、また日本語の形態素解析 [9] と後述する地名情報を利用してそのウェブページが含む地名のリストを生成した。

地名の辞書には国土地理院刊行の「数値地図 2500」[10] から、市区町村名・駅名、学校などの公共施設の名称と代表点の座標のリストを抽出した。また

数値地図にはよくランドマークとして利用される地理データが不足しているので、次のようにして「交差点」と「寺社仏閣」のデータを生成し、上記のリストに加えた。交差点に関しては、京都独自の慣習☆を利用し、数値地図のもつ道路ネットワーク情報から自動的に交差点の名称とその座標のリストを生成した。寺社仏閣に関しては、寺社仏閣の住所リストを利用し、数値地図から生成した市区町村名を参照しながら地図上にマッピングした。なお、すべての地名データは数値地図の情報を元にして、“左京区が京都市に含まれる” といった地名の階層構造を復元してある。

4.2 地域における人気度の計算実験

2 で説明した各計算式によって、ウェブページの人気度を計算し、その結果を比較することで、対象を限定するウェブページの人気度計算の有効性を示す実験を行う。なお、それぞれの実験は京都市内のいくつかの小地域に対して行うものとする。

4.2.1 対象限定による人気度の変化

次の 2 通りの人気度によるランキング手法を比較する実験を行った。対象地域は、京都市内の代表的な観光地である “銀閣寺”、および代表的な繁華街である “四条河原町” を選択した。

グローバルな人気度 収集した約 120,000 の全ウェブページを対象にして、式 (2) によって計算した従来の手法による人気度。ただし、α は、Page 氏らの手法に習い $\alpha = 0.87$ として計算している。

ローカルの人気度 特定の地域の地名 l を含むウェブページ集合 L に対して、式 (3) を適用した単純な対象限定による人気度

各ページには “充実した内容を持つ重要なページかどうか”、“対象地域の情報として適切か” などの基準により、著者の主観で 0, 1, 2, 3, 4, 5 の 6 段階の点数をつけた。上位 n 件のウェブページの平均点で結果を評価した結果を図 4 に示す。グラフの横軸 n に対し、縦軸は上位 n 件のウェブページの平均点である。

グローバルな人気度では、内容が充実していても地域情報を持たないページが多く上位にランクされ、全般的に評価が低い。逆にローカルな人気度では、地域情報を持つ内容が充実したページが上位にくることが多く、全般的に評価が上がっている。

また繁華街である “四条河原町” での結果が全般的に悪かった。これは交通の要所であるために、別の地域情報のウェブページの中で “四条河原町からのアクセス” などの表現が頻出することが原因と思われる。

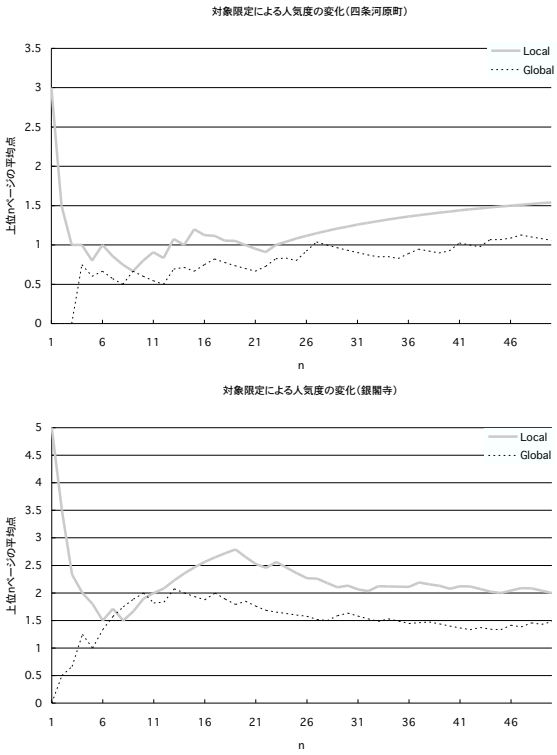

図 4 対象限定による人気度の変化

4.2.2 地域拡張による人気度の変化

人気度計算の対象となる部分グラフを地域によって拡張することにより、ウェブページの人気度がどのように変化するのかを調べる実験を行った。

対象地域から 500m,1km,2km 以内のいずれかの地名を含むウェブページ集合 L_{500m}、L_{1km}、L_{2km} を作成し、それぞれに式 (3) を適用して人気度計算を行い、4.2.1 と同様に著者の主観によるウェブページの評価との比較を行った。この結果を示したのが図 5 である。

500km、および 1km の拡張によって、全般的に上位に地域情報を持たないウェブページがくることが減り、やや結果が向上している。これより、地域にある程度幅を持たせることにより、地域の人気度がより適切に求められることがわかる。

しかし、2km の拡張では逆に地域と無関係なウェブページが上位に現れることが多くなっていることがわかる。これは、対象地域を広げすぎたために部分グラフを構成するウェブページの中に、地域と無関係なページがかなり増えてしまったため、この地域と無関係なページからのリンクが計算結果に及ぼす影響が無

☆ 京都では、場所を表すのに住所表記ではなく交差点をランドマークとして利用するのが一般的である。また、交差点の名称は一部の例外を除き、交差する通りの名前の組み合わせで生成される。例えば烏丸通と四条通の交差点は “四条烏丸” または “烏丸四条” と呼ばれる。

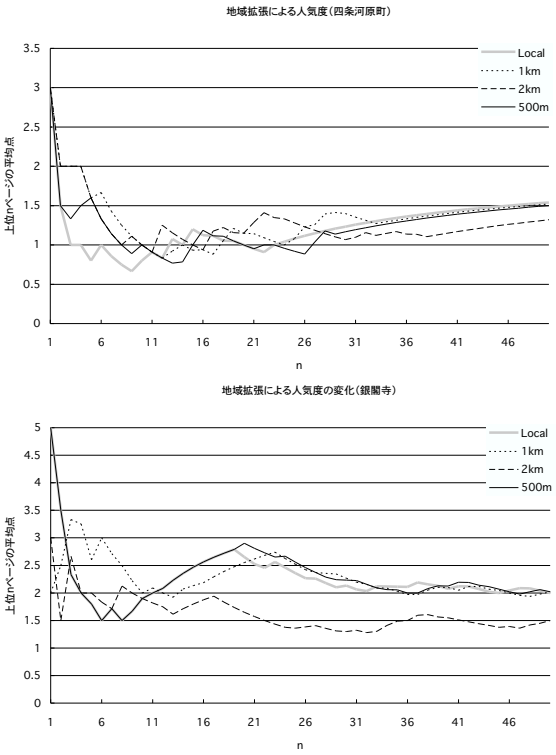

図 5 地域拡張による人気度の変化

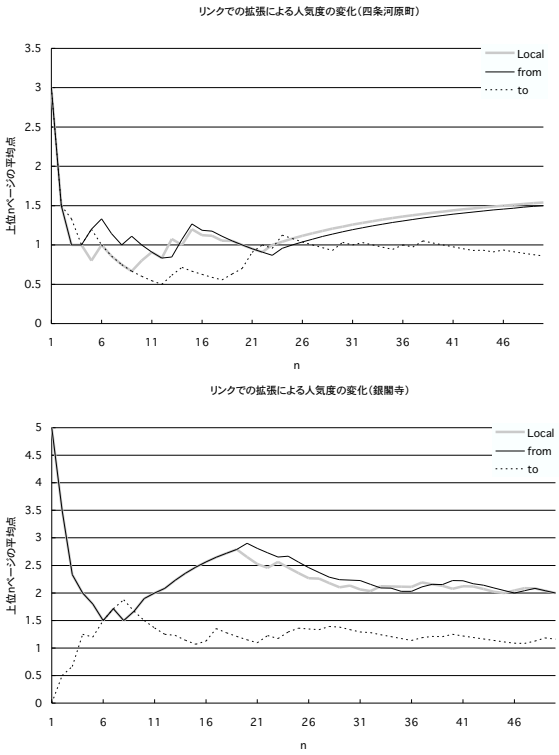

図 6 リンクでの拡張による人気度の変化

視できないほど大きくなったと推定される。限度を超えて地域を拡張することで、地域情報を得にくくなってしまうといえるだろう。

繁華街の情報のほうが、地域拡張による結果の向上が確実に現れている。これは、地理的に近い地名をもつページとのリンクを利用することで、前述のような“四条河原町からのアクセス” などの表現をもつウェブページからのリンクの重みが相対的に下がったことによると思われる。

4.2.3 リンクでの拡張による人気度の変化

リンクによって対象地域を拡張することにより、ウェブページの人気度がどのように変化するのか調べる実験を行った。

まず、リンクを利用して、L のページに対してリンクをしているウェブページの集合 L_{from}、L のページからリンクされているウェブページの集合 L_{to} を抽出した。

次に、$L \cup L_{from}$ および $L \cup L_{to}$ のような拡張部分グラフを作成し、4.2.1、4.2.2 と同様にして拡張部分グラフに対してウェブページの人気度を計算した。この結果を著者の主観の評価と比較した結果を示したのが図 6 である。

特徴的な傾向として、$L_{from} \cup L$ によって部分集合の構成要素がほとんど増えていないことがあげられる。母集団がほぼ同じであるため、結果もほぼ同じであるが $L_{from} \cup L$ の方が若干よい結果がでている。

逆に、$L_{to} \cup L$ の拡張では明らかに誤差が大きかった。人気度計算は多くの場合において、被リンク数の影響が大きく、このリンク元を集合に追加する拡張では、被リンク数が全ページを対象にした人気度計算の場合と等しいため、地域を限定したことによる影響がほとんどなくなってしまっていることがわかる。

4.3 地域指向性の計算

4.3.1 実 験 手 順

3 で示したウェブページの地域指向性の計算手法の有効性を示す実験を行う。実験 1 と同様の “銀閣寺”、“四条河原町” の 2 地域とやはり京都市内の観光地である “祇園” を対象に、各地名を含むそれぞれのウェブページ集合に対し、式 (4) の計算を適用し、地域指向性の計算を行った。

計算結果は人手によって地域情報を持つかどうかを判断したデータと比較して適合率と再現率を計算し、それによって評価する。ただし、システムが判断する正解の条件は $b(w_j) > 2.0$ とした。実験の結果を図 7 に示す。

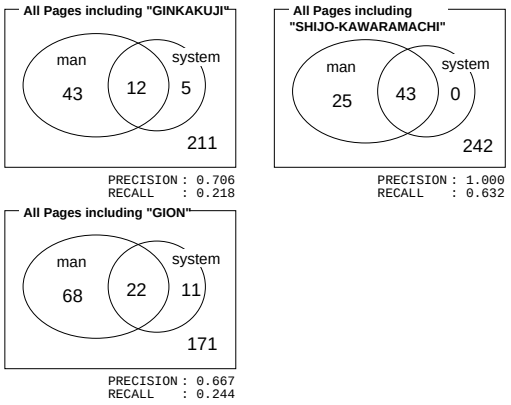

図 7 各地域のウェブページにおける地域指向性の計算結果

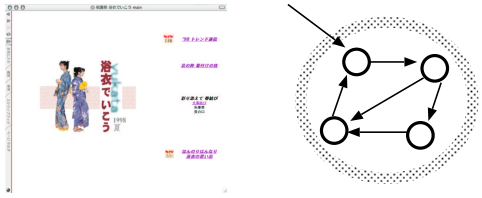

図 8 地域情報をもつと判断されたが実際は持たないページ例とその周辺のリンク構造

4.3.2 結果と考察

適合率については、観光地の情報、繁華街の情報の違いを問わず高い値を得た。

しかし、地域と無関係なウェブページが地名を含み、かつ相互リンクしている場合に、これらの地域に無関係な情報がシステムによって地域情報として認識されてしまう。この問題はおもに同一サイト内での相互リンクによって引き起こされ、例として図 8 のような場合があげられる。

図 8 のウェブページは、「祇園祭 浴衣でいこう」というタイトルのページであるが、実際には「浴衣」を話題にしたウェブページであり、地域情報を提供するものではない。しかし、図 8 の右側のリンク構造で示されるように、サイト内に同じタイトルをもつページが強連結構造をもち、かつ外部へのリンクを持たないため、結果として式 (4) の計算において高い値を出していることがわかる。銀閣寺のデータでの不正解例にも同様の傾向が見られた。

適合率と比較して、再現率はよい結果がでていない。ウェブページが地域情報を含むかどうかを、同一地域の地名を含むウェブページにリンクをしているかどうかで判断しているため、リンクを 1 つも持たないウェブページがすべて棄却されていることが主要因であった。謝ってローカルでないと判断されたウェブページのうち、リンクを一切持たないページの割合が非常に高く、これらを拾い上げることが今後の課題となる。

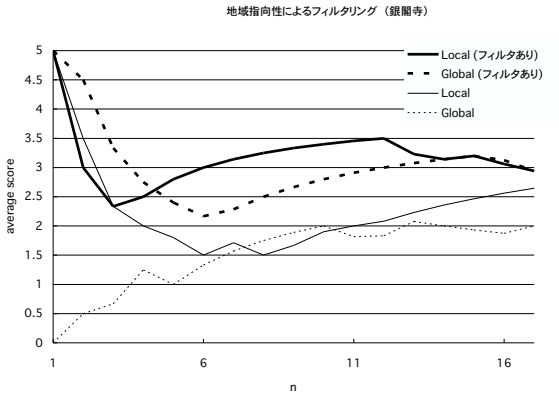

図 9 地域指向性によるフィルタリング

4.4 地域での人気度と地域指向性によるフィルタリング

4.3 の地域指向性の計算により、ローカルな情報であると判断されたウェブページだけを選択し、4.2 で計算した地域を限定した人気度によってランキングし、その結果を 4.2 と同様に著者の主観によるデータとの比較により評価した。実験は銀閣寺のデータに対して行った。その結果を図 4.4 に示す。縦軸は、上位 n のウェブページの平均点である。

図 4.4 より、ローカル・グローバルな人気度双方において計算の精度が向上していることがわかる。特にグローバルな人気度計算での結果の向上が著しい。地域指向性によるフィルタリングを併用することで、地域と無関係なウェブページがランクからはずれるためと思われる。

5. 地域での人気度と地域指向性の融合

4 章での実験により、地域での人気度および地域指向性の双方の高いウェブページを選択することで、効果的に地域情報を持つウェブページを抽出できることを示した。

しかし、地域指向性においてはリンク先に地域情報があることが前提であるため、リンクをもたないウェブページが排除されてしまうという問題点がある。そのため地域指向性が高いページに対象を限定することで多くの有用な情報を除外してしまっていると考えられる。

そこで、地域指向性の結果を地域での人気度計算の中に有機的に結びつけることでこの問題を回避する方法を提案する。

地域での人気度計算の中で、リンク先のページの地域指向性の値によって遷移確率に重みをつける。すなわち、ユーザはより地域指向性の高いウェブページへ誘導されるという前提で人気度計算を行う。

6. 関連研究

Junyan Ding 氏らの研究 [4] で、ウェブページの地理的な有効範囲の計算手法を提案するものがある。この手法は、ウェブページの存在するホストなどの情報を利用してウェブページを地図上にマッピングしたのち、ウェブページのリンクの分布や、ウェブページのテキスト内容を利用して有効範囲を特定している。あるウェブページ w について、ある地域のウェブページのうち w へのリンクを含むものの割合は、なだらかに分布しているという前提に基づき、リンクの分布から地理的な有効範囲を求めることができる。

京都大学の田中克己氏らの研究 [8] で、ウェブページのコンテンツに含まれる地理情報と、コンテンツの話題からローカル度を測定する研究がある。この研究では、ウェブページに含まれる地名が特定の狭い地域に集中しているかどうかでローカル度を判断する。また、ローカルなウェブページは日常的な話題をあつかったものが多いという考えに基づき、類似度の高いページが多いページほどローカルと見なしている。

本論文ではウェブのリンク構造を利用しているのに対し、この研究では１つの文書の内容と他の関連する類似するページからローカル度を検出している点が異なる。

7. 結　　論

ウェブページの人気度はウェブ情報検索によって得られたウェブページを評価する上で非常に有効な指標である。本論文では、この人気度計算手法を地理情報システムの持つ地域知識と統合して用いることで、特定の地域における人気度を定義した。また、地域における人気度は地域の限定による部分グラフに既存の方法を適用するだけでは不十分であることを示し、地域での人気度を計算する上で効果的な改良法について述べた。

また、ウェブページと地域との関係は、ウェブページの地域における人気度だけではない。リンク先のウェブページの内容とリンクを評価することで、ウェブページから地域に向けられる指向性も評価できるようになった。

これらの手法を統合して用いることにより、地域情報を含むだけでなく、地域から高い評価を受けた、地域に密着した情報を抽出することができる。

地域情報検索の精度が上がれば、ウェブ上における地域のコミュニケーションを促進できる。また、ウェブページと地域との関係を評価することができるので、ウェブセマンティックスの観点からも地域知識の発見に本研究を役立てることも可能であろう。

しかしながら、現在の手法のままでは、ウェブ情報検索の度に多くの計算を必要とし、検索の計算コストが非常に大きくなるなど問題点も多い。今後は、実装・実験を通して検索精度のさらなる向上を計るだけでなく、検索の度に再計算をしなくてもすむような近似計算手法などについても検討する予定である。

謝辞　本研究は科学技術振興事業団（JST）・戦略的基礎研究推進事業（CREST）における「デジタルシティのユニバーサルデザイン」プロジェクトの支援によって行われました。ここに記して謝意を表すものとします。

参 考 文 献

1) Ryong Lee, Hiroki Takakura, Yahiko Kambayashi: “Visual Query Processing for GISs with Web Contents”, VDB6 (May 2002) (to appear);
2) 井上陽介　李龍　高倉弘喜　上林弥彦: “地域情報検索のための対象を限定したウェブページの人気度”, DBWeb 2001 (Dec. 2001).
3) K.S. McCurley: “Geospatial Mapping and Navigation of the Web,” WWW10, (2000).
4) J. Ding, L. Gravano and N. Shivakumar: “Computing Geographical Scopes of Web Resources”, VLDB2000, pp.545-556, (2000).
5) S. Brin, and L. Page : “The Anatomy of a Large-Scale Hypertextual Web Search Engine”, Computer Networks and ISDN Systems, Vol. 30, pp. 1-7, (1998).
6) Masatoshi Arikawa, Koji Okamura: “Spatial Media Fusion Project”, Proc. of Kyoto International Conference on Digital Libraries: Research and Practice, pp.75-82, (Nov. 2000).
7) X. Zhou, J.D. Yates and Guihai Chen: “Searching the Web Using a Map,” Internation Conference on Web Information Systems Engineering. Vol. 1, pp. 117-124, (2000).
8) 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”, DBWeb2001 (Dec. 2001).
9) Morphological Analyzer Chasen http://chasen.aist-nara.ac.jp/
10) 国土地理院: 数値地図 2500 http://www.jmc.go.jp/