空間情報ハブ抽出のためのウェブリンク解析手法について

SD-1-4

Web Link Analysis for Extracting Spatial Information Hub Pages

張 建偉[1] 石川 佳治[2] 北川 博之[2]
Jian Wei Zhang Yoshiharu Ishikawa Hiroyuki Kitagawa

筑波大学システム情報工学研究科[1]
Graduate School of Systems and Information Engineering
University of Tsukuba

筑波大学電子・情報工学系[2]
Institute of Information Sciences and Electronics
University of Tsukuba

1 まえがき

ウェブの爆発的な拡大により，大量のウェブページの中から有用な情報を抽出する技術はより重要さを増している．そのための手法として，近年ウェブマイニング[1, 2]の研究が盛んに進められている．特に，ウェブページ間のリンク情報を用いるリンク解析は，評判の高いウェブページを特定するための重要な技術となっている．

一方で，携帯機器やGPSの普及などにより，位置に応じて適切な情報提供を行うためのサービスが現在重要となってきている．そのようなサービスの1つとして，指定された地点の周辺情報に関するウェブページ群をユーザに提示する，地域性を考慮したサーチエンジンや，特定の地域に関するウェブ情報を提供する研究が進められている[3, 4, 6, 7, 5, 8]．これらの研究では，ウェブページのコンテンツ解析やリンク解析の手法を用いて，地域性を有するウェブページの特定などを行っている．

これらの関連研究の動向を踏まえ，本研究ではウェブページ群からの**空間情報ハブ**の抽出手法の開発を行う．本研究では空間情報ハブを，「ある地域に関するウェブページ群に関して有用な多くリンクを張っている定評あるウェブページ」と定義する．ウェブ上からこのようなページを抽出することができれば，ある地域に関するポータルサイトとして活用することが可能となる．

2 関連研究

2.1 ウェブマイニング

ウェブマイニング[1, 2]は，大別すると，1) ウェブページのコンテンツマイニング，2) リンク解析，3) ウェブサイトのログ解析の3つのアプローチに分けることができる．本研究は，2) のリンク解析を中心に空間情報ハブの抽出を図る．リンク解析手法の代表例としては，Googleで用いられているPageRank[9]や，ユーザが指定したトピックに関してハブとオーソリティのページを抽出するHITS[10]が挙げられる．

ここでは特に本研究で拡張を図るHITSについて簡単に説明する．まず，前もってユーザが指定したキーワードにより数百ページ程度のウェブページをサーチエンジンなどで抽出してルートセットとする．次に，ルートセット内のページからリンクされているページの集合とルートセット内のページをリンクしているページの集合をサーチエンジンなどを利用して求める．これらのページ群から構成されるウェブの部分グラフを V とする．

HITSのアルゴリズムを図1に示す．各ページのハブ度（そのページが良いオーソリティのページをリンクしている指標）をベクトル $\boldsymbol{a}$ で，オーソリティ度（そのページが良いハブからリンクされている指標）をベクトル $\boldsymbol{h}$ で表す．一様な値に設定した初期値から，5〜8行目の繰返し処理と9〜10行目の正規化処理により，スコアが収束するまで繰り返したときベクトル $\boldsymbol{a}_t, \boldsymbol{h}_t$ には，それぞれオーソリティ度，ハブ度の計算結果が入る．

```
1  1 := [1, ..., 1] ∈ R^|V|
2  a_0 := h_0 := 1                          // スコアを初期化
3  t := 1
4  repeat
5    foreach v ∈ V do
6      a_t(v) := Σ_{w∈parent[v]} h_{t-1}(w)  // オーソリティ度の更新
7      h_t(v) := Σ_{w∈child[v]} a_{t-1}(w)   // ハブ度の更新
8    end
9    a_t := a_t/||a_t||                     // スコアの正規化
10   h_t := h_t/||h_t||
11   t := t + 1
12 until ||a_t − a_{t-1}|| + ||h_t − h_{t-1}||   // スコアが収束するまで
13 return (a_t, h_t)
```

図1 HITSのアルゴリズム

2.2 地域性を考慮したウェブ情報の収集・探索

ウェブの中から特定の地域に関するページを抽出するための研究としてさまざまな手法が提案されている．[3]では，位置情報をウェブから収集する手法について述べている．[4, 5]では，ウェブページ内に含まれる地名・組織名などの地理情報，ページ内の話題の偏在性，話題の注目度など，さまざまな要素を考慮して，ページのローカル度を与える手法を提案している．ローカル度は，そのページが地域密着型の情報を有しているかの判断に利用する．ローカル度のアプローチとは異なるが，本研究においてもウェブページがどの程度地域密着型の情報を表しているかを，間接的にリンク解析に反映している．

KyotoSEARCH[6, 7]では，京都を対象に，効率的に特定地域に関する情報検索を支援するシステムを開発している．特に[6]では，地域を限定したページ集合に対して地域性を考慮してリンク解析するためのPageRankの拡張手法を提案している．一方，[8]では，地域情報サービスを提供するため，ウェブ空間を拡張するアプローチを述べている．通常のウェブページ間のリンク以外に地理空間上へのリンクを用いてウェブを拡張することで，地理空間を経由したウェブ空間のナビゲーションが可能となる．一方，本研究では，地理的な実空間へのリンク

の概念をウェブマイニングのために利用する．

3 提案手法の概要

3.1 ウェブページ群からの空間情報の抽出

前処理として，ウェブデータの収集を行い，収集した各ウェブページの中から，住所，郵便番号，施設名などの抽出を行い，その座標値を計算する．空間情報の抽出においては，正確性を重視して，正確な座標が特定できるようなフル表記の住所や7桁の郵便番号などの情報を利用する．これにより，各ウェブページには一般に複数個の座標値が対応することになる．

3.2 ベースセットの構築

リンク解析処理は，ユーザから分析対象の地理領域が指定された時点で開始する．指定された地理領域に対し，その中に関連する座標値の少なくとも1つが含まれるようなページを収集し，これをルートセットとする．HITSと同様，このルートセットのページに関連するページを追加し，ベースセットを構築する．すなわち，図2の左側のような，指定された地理領域に関連するページ群からなるウェブ空間のサブグラフを構築する．

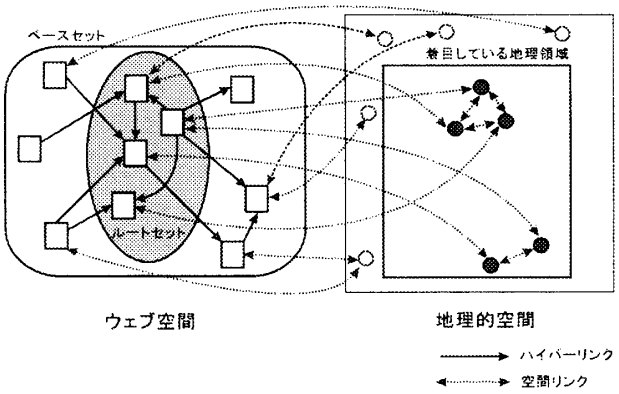

図2 提案手法のアプローチ

HITS ではユーザ指定のキーワードをもとにルートセットを構築するのに対し，本手法では空間的な条件をもとに構築する点に大きな違いがある．

3.3 空間ノード・空間リンクの生成

次に，空間ノードと空間リンクを生成する．空間ノードとは，ベースセット内のウェブページ中に含まれる住所・郵便番号などの地理的情報に対応するノードである．ただし，ノードは指定された地理領域内の地理的情報についてのみ生成するものとし，同じ表記（例：同一の郵便番号や住所表記）については1つだけ生成する．

次いで，空間ノードとその情報を含んでいたページの間に双方向のリンクを作成する．このリンクは，ウェブのハイパーリンクと異なり，空間的な関連性に基づくことから，空間リンクと呼ぶ．また，2つの空間ノードの間の距離がある閾値以下である場合にも，お互いが近傍にあるものと考え，双方向の空間リンクを生成する．

3.4 リンク解析処理

構築されたグラフは，ウェブ空間上における近さ（意味的な関連や組織・社会的な関連を反映）と地理空間上の近さを融合したものとなっている．このグラフに HITS アルゴリズムを適用し，ウェブ空間側のノードのハブ度とオーソリティ度を求めることにより，指定された地理空間上におけるウェブページの評価値を与える．

複数の空間情報を含むウェブページの場合，それらの地理的位置が離れていたり，着目している領域外の空間情報が含まれる場合も多い．本研究では，着目領域内への空間リンクを多数有しているページの重要性を高め，他の領域の情報を多く含むページの重要性を低くするために，リンクに重み付けを行うリンク解析手法 [11] の拡張を検討している．

4 まとめと今後の課題

本稿では，特定の地理領域に関し有用なページをリンクする良質のウェブページである空間ハブを，ウェブのリンク解析によって求める手法の概略を述べた．今後は提案手法の具体化と実験を行う予定である．

謝辞

本研究の一部は，日本学術振興会科学研究費若手研究(B)(14780316)，基盤研究(B)(12480067)，および文部科学省科学研究費特定領域研究 (14019009) による．

参考文献

[1] S. Chakrabarti, *Mining the Web: Analysis of Hypertext and Semi Structured Data*, Morgan Kaufmann, 2002.

[2] P. Baldi, P. Frasconi, and P. Smyth, *Modeling the Internet and the Web: Probabilistic Methods and Algorithms*, Wiley, 2003.

[3] 横路誠司, 高橋克巳, 三浦信幸, 島健一, 位置指向の情報の収集, 構造化および検索手法, 情報処理学会論文誌, Vol. 41, No. 7, pp. 1987–1998 (2000).

[4] 馬強, 松本 知弥子, 田中克己, 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, 情報処理学会研究報告, DBS-128-69, pp. 515–522 (2002).

[5] C. Matsumoto, Q. Ma, and K. Tanaka, Web Information Retrieval Based on the Localness Degree, *Proc. DEXA 2002*, LNCS 2453, pp. 172–181 (2002).

[6] 井上陽介, 李龍, 高倉弘喜, 上林弥彦, 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, 電子情報通信学会データ工学ワークショップ (2002).

[7] 李龍, 椎名宏徳, 高倉弘喜, 上林弥彦, 地域ウェブ情報検索のための2次元領域質問処理法, 電子情報通信学会研究報告, DE2003-61 (2003).

[8] 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. SIG 6(TOD 7), pp. 81–90 (2000).

[9] 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).

[10] J.M. Kleinberg, Authoritative Sources in a Hyperlinked Environment. *JACM*, Vol. 46, No. 5, pp. 604–632 (1999).

[11] K. Bharat and M.R. Henzinger, Improved Algorithms for Topic Distillation in a Hyperlinked Environment, *Proc. SIGIR*, pp. 104–111 (1998).