2004－DBS－134　(II)　(55)
2004／7／14

空間情報ハブ抽出のためのウェブリンク解析手法の開発

張建偉†,　　石川佳治†,　　北川博之†

ウェブの爆発的な拡大から，大量のウェブページの中から有用な情報を抽出するためのウェブマイニングの研究が盛んに進められている。一方では，携帯機器や GPS の普及から，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。本研究では，これらの動向を踏まえ，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。本手法では，地理的な情報をもとに拡張されたウェブのリンク構造を解析することにより、着目している地域におけるページの評判や有用性を判定する。ウェブのグラフ構造を，着目している地理領域に関連した空間ノードと空間リンクにより拡張する点が特徴となっている。

Web Link Analysis for Extracting Spatial Information Hub Pages

Jianwei ZHANG†,　Yoshiharu ISHIKAWA†,　Hiroyuki KITAGAWA†

Recently web mining which tries to find relevant information from the vast amount of web pages has attracted a lot of research interests. On the other hand, it is becoming an important task to provide information related to a certain geographical area. In this paper, we propose an algorithm to extract spatial information hub pages. A spatial information hub is a webpage about a certain geographical area, which has much geographical information or many hyperlinks . We employ geographical information to create spatial nodes and spatial links, and then conduct link analysis based on this extended link structures.

1　はじめに

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っている [12]．本手法の特徴の一つは，ハイパーリンクで構成されるウェブのグラフ構造を，着目している地理領域に関連した**空間ノード**と**空間リンク**により拡張する点にある．従来のウェブのリンク解析がウェブのリンク情報をもとにページの評判や有用性を判定していたのに対し，本手法では地理的な情報をもとに拡張されたウェブのリンク構造を解析する．これにより，着目している地域における評判や有用性という面も考慮してページを評価する．このような意味で，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や地理情報に関して有用な多くのリンク（ハイパーリンクおよび空間リンク）を張っているウェブページ

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，

†筑波大学大学院
システム情報工学研究科コンピュータサイエンス専攻
Department of Computer Science, Graduate School of Systems and Information Engineering,
University of Tsukuba

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．本研究では，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** (Hypertext Induced Text Search) [10] 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 $\boldsymbol{h}$ 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 $\boldsymbol{a}$ 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 $\mathbf{1} := [1, \ldots, 1] \in \mathcal{R}^{|V|}$ // スコアを初期化
2 $\boldsymbol{a}_0 := \boldsymbol{h}_0 := \mathbf{1}$
3 $t := 1$
4 **repeat**
5 　**foreach** $v \in V$ **do** // オーソリティ度の更新
6 　　$\boldsymbol{a}_t(v) := \sum_{w \in parent[v]} \boldsymbol{h}_{t-1}(w)$ // ハブ度の更新
7 　　$\boldsymbol{h}_t(v) := \sum_{w \in child[v]} \boldsymbol{a}_{t-1}(w)$
8 　**end** // スコアの正規化
9 　$\boldsymbol{a}_t := \boldsymbol{a}_t / \|\boldsymbol{a}_t\|$
10 　$\boldsymbol{h}_t := \boldsymbol{h}_t / \|\boldsymbol{h}_t\|$
11 　$t := t + 1$ // スコアが収束するまで
12 **until** $\|\boldsymbol{a}_t - \boldsymbol{a}_{t-1}\| + \|\boldsymbol{h}_t - \boldsymbol{h}_{t-1}\|$
13 **return** $(\boldsymbol{a}_t, \boldsymbol{h}_t)$

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という**ローカル性**を考慮している．ローカル性が高いページやローカル性が高いページを多数リンクしているページをより高く評価するようにリンク解析処理を工夫している．

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，着目している地理情報を表す空間ノードと，空間ノード間および空間ノードとウェブページ間の空間リンクを導入することで，ウェブ空間を拡張する．[8] が拡張したウェブ空間をユーザに提供するナビゲーション機能に利用していたのに対し，本研究ではリンク解析によるウェブページからの情報抽出・知識発見に用いる点が大きく異なる．

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．実際のウェブページには，これら以外にも地理情報に関連付け可能な多数の情報が含まれている．しかし，部分的な地名（例：「東京」

「春日」）などは抽出が困難であいまい性が高いため，ウェブページ内に含まれる地理情報に対応する地点を正確に特定することが難しいため，抽出は行わない．以上の処理により，各ウェブページには一般に 0 個以上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，ルートセットに含まれるページへのリンクを有するウェブページ，および，ルートセットに含まれるページがリンクしているウェブページの追加を行う．これにより，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．

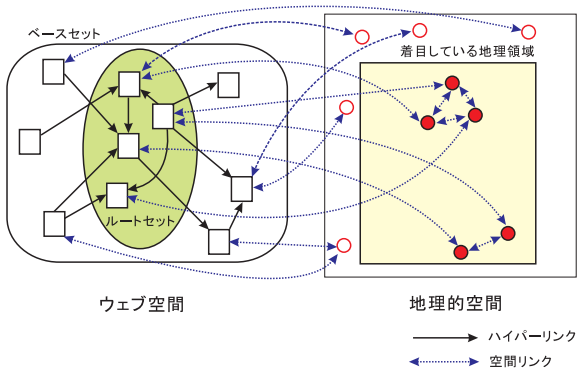

図 2: 拡張ベースセットの構築

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，同じ表記（例：同一の郵便番号や住所表記）については空間ノードを 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．

本手法では，空間ノードと空間リンクを前節で説明したベースセットに追加することにより，ベースセットを拡張する．以下ではこれを**拡張ベースセット**と呼ぶ．あいまい性が生じない場合には，そのままベースセットと呼ぶ．これにより，たとえば 2 つのウェブページの間にハイパーリンクのつながりがない場合でも，位置が近い地理情報を含んでいる場合などにはお互いの間に関連性が生じることになる．このように拡張されたベースセットをもとにリンク解析を行うことで，本手法では地理的な情報も考慮してウェブページの評判や有用性を評価する．

3.4 リンク解析処理

前節のように構築された拡張ベースセットのグラフ構造は，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．

3.4.1 リンクの入出次数に基づく指標の導入

本手法では，図 2 のように，各ウェブページ内に現れるすべての空間情報に対応して空間ノードが存在し、そこへの空間リンクがあると考える．しかし，実際のリンク解析では，着目している地理領域内の空間ノード，および，それらへの空間リンクのみを分析対象とする．つまり，リンク解析の際には，仮想的に存在する空間ノード，空間リンクのうち，着目している地理領域に関連するもののみを実際に考慮する．ただし，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ため，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，ノードやリンクに重み付けを行うリンク解析手法 [11] の考え方を拡張し，以下の 2 つの指標を定義する．

ノード v から出るリンク（ハイパーリンクおよび空間リンク）の総数を $outlinks(v)$ とし，そのうち拡張ベースセット内のノードを指すリンクの数を $effective\_outlinks(v)$ とする．このとき，$out\_ratio(v)$

を，

$$out\_ratio(v) = \frac{effective\_outlinks(v)+1}{outlinks(v)+1} \quad (1)$$

と定義する．この値は，含んでいるリンクが拡張ベースセット内のノードを指すほど高くなり，拡張ベースセット外のノード（地理的な意味でに関連が低いノード）を指すほど低くなるという性質がある．

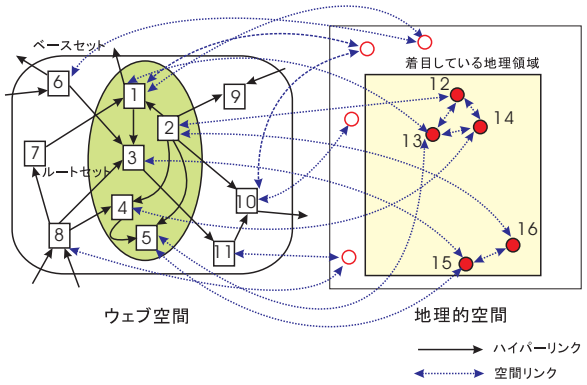

図 3: 分析対象のベースセットの例

たとえば，図 3 において，ノード 1 から出るリンクは 5 本があるが，そのうち拡張セットベース内へのリンクは 2 本である．よって，ノード 1 については，

$$out\_ratio(1) = \frac{2+1}{5+1} = \frac{1}{2} \quad (2)$$

となる．

同様に，ノード v に入るリンク（ただし，ここでは収集さたウェブページ集合内に含まれるリンクのみを考える）のうち，ノード v へ入るリンクの総数を $inlinks(v)$ とする．そのうち，拡張ベースセット内のノードからのリンク数を $effective\_inlinks(v)$ とする．これらをもとに，$in\_ratio(v)$ を，

$$in\_ratio(v) = \frac{effective\_inlinks(v)+1}{inlinks(v)+1} \quad (3)$$

と定義する．図 3 のノード 1 に関しては，

$$in\_ratio(1) = \frac{3+1}{5+1} = \frac{2}{3} \quad (4)$$

となる．

3.4.2 拡張した HITS によるリンク解析

本手法では，前節で提案した out_ratio, in_ratio を HITS のアルゴリズムに導入する．これにより，拡張ベースセット内のノードよりも外のノードと関連が深いノードのハブ度，オーソリティ度のスコアをより低く評価することができ，逆に，拡張ベースセット内のノードと密接に関連しているノードの評価を高くできると考えられる．

拡張した HITS のアルゴリズムを図 4 に示す．6 行目と 7 行目で上記の指標による重み付けを行っている点が相違点である．

1 $\mathbf{1} := [1, \ldots, 1] \in \mathcal{R}^{|V|}$
2 $\boldsymbol{a}_0 := \boldsymbol{h}_0 := \mathbf{1}$
3 $t := 1$
4 **repeat**
5 **foreach** $v \in V$ **do**
6 $\boldsymbol{a}_t(v) := in\_ratio(v) \times \sum_{w \in parent[v]} \boldsymbol{h}_{t-1}(w)$
7 $\boldsymbol{h}_t(v) := out\_ratio(v) \times \sum_{w \in child[v]} \boldsymbol{a}_{t-1}(w)$
8 **end**
9 $\boldsymbol{a}_t := \boldsymbol{a}_t / \|\boldsymbol{a}_t\|$
10 $\boldsymbol{h}_t := \boldsymbol{h}_t / \|\boldsymbol{h}_t\|$
11 $t := t + 1$
12 **until** $\|\boldsymbol{a}_t - \boldsymbol{a}_{t-1}\| + \|\boldsymbol{h}_t - \boldsymbol{h}_{t-1}\|$
13 **return** $(\boldsymbol{a}_t, \boldsymbol{h}_t)$

図 4: 拡張した HITS のアルゴリズム

3.4.3 スコアの計算例

表 1: 提案手法によるスコアの計算結果

ノード番号	ハブ度（×100）	オーソリティ度（×100）
1	11	22
2	55	16
3	21	20
4	29	32
5	26	51
6	4	0
7	8	2
8	14	0
9	0	12
10	0	11
11	3	27
12	31	7
13	39	40
14	28	31
15	32	32
16	17	21

図 3 の拡張ベースセットに対するハブ度，オーソリティ度の計算結果を表 1 に示す．ハブ度が最も高いノードはノード 2 で，ノード 5, 3 がそれに続いている．ウェブページとして見た場合，ノード 5 は外向きのリンクがなくハブとはなり得ないが，本手法では良質の空間情報を有している（良質の空間リンクを多数有している）と評価されて，高いハブ度となっている．

比較のため，従来の HITS アルゴリズム（図 1）により，拡張前のベースセット（3.2 の処理で生成されたグラフ）を分析した場合のスコアを表 2 に示す．ハ

表 2: 従来の HITS による計算結果

ノード番号	ハブ度（×100）	オーソリティ度（×100）
1	9	40
2	87	0
3	14	21
4	17	47
5	0	40
6	9	0
7	17	13
8	35	0
9	0	33
10	0	40
11	17	34

ブ度について見た場合，ノード 2 が高いスコアとなっている点は提案手法と同様であるが，ノード 5 の評価が低く，逆にノード 8 の評価が高い点などが異なる．提案手法では，ノード 8 は空間情報との直接的な結びつきがないため低く評価されているが，従来の HITS では高く評価している．

4 実験

4.1 実験の準備

実験に利用するのは国立情報学研究所により提供されている NTCIR-4 WEB タスク文書データ（NW100G-01）である．これは主として.jp ドメインから 2001 年に収集された HTML もしくはプレーンテキストファイルからなる．97,561 個のウェブサイトから文書 11,038,720 件を収集し，79,699,256 個のリンクが含まれている．

前処理として，NW100G-01 データから各ページのメタ情報（URL など），ページの間のリンク情報などをデータベースに追加し，リンク解析を効率的に処理可能とするための connectivity server を構築した．空間情報に関しては，7 桁の郵便番号を中心に情報の抽出と空間的な位置との対応付けを行い，connectivity server に登録を行った．具体的には 86,315 個の郵便番号を抽出している．

4.2 実験結果

4.2.1 in/out_ratio 導入に関する比較

第一の実験として，空間ノード間に空間リンクを張るかどうかの距離の閾値を 0.002 に限定し，着目領域を東京都豊島区池袋本町（〒 1700011）から距離が 0.015 の地域に指定する．なお，ここでの距離は，緯度，経度を x 座標，y 座標と近似的にみなした座標値をもとに計算する．これにより，空間ノードが 157 個，ウェブページと空間ノードの間の空間リンクが 296 個，空間ノードの間の空間リンクが 896 個生成された．

提案手法によるハブ度の上位 6 件の実験結果を表 3 に示す．各ページのリンク情報を表 4 に示す．doc_id は文書の ID である．spa_link は文書中の全ての空間リンク数で，$effec\_spa\_link$ はそのうち着目している地域内の空間ノードへの空間リンク数である．$weblink$ は文書中の全てのウェブリンク数で，$effec\_weblink$ はそのうちベースセット内のウェブページへのウェブリンク数である．out_ratio は全てのリンク（ウェブリンクと空間リンク）のうち有効なリンクの割合である．$evaluation$ は各ページに対して，良いハブページかどうかの評価である．なお，各ページの内容とリンク先のページの内容を見ることで評価した．○は良いハブページを表し，×は良くないハブページを表す．△のページはどちらとも言えないものである．

比較のため，空間ノードと空間リンクを生成するが，in/out_ratio の指標を導入しない手法による実験結果を表 5 に示す．各ページのリンク情報を表 6 に示す．

表 6 で示されたように，in/out_ratio の指標を導入しないと，ハブ度が上位になるページはウェブリンクと空間リンクを多く持つが，そのうち着目している地域内への空間リンクの割合が低い．即ち，着目している地域以外の空間情報を多く含むという欠点がある．ハブ度が 1 位になるページ 10822643 は東京 23 区のダイビングショップとサービスのページである，このページは東京 23 区の空間情報（郵便番号）を 62 個含んでいるが，そのうち指定された地域（豊島区池袋本町周辺）内の空間情報は 4 個しかないため，良いハブページと見なされない．

表 4 で示されたように，in/out_ratio を導入することで，このようなページはより低く評価される．提案手法でハブ度が 1 位になるページ 13039139 は東京都豊島区の旅行代理店のリストである．このページは 8 個の空間情報を含んでいる．そのうち指定された地域内の空間情報は 6 個あるため，良いハブページと見なされる．

4.2.2 閾値の影響

第二の実験として，空間ノード間に空間リンクを張るかどうかの距離の閾値を 0.005 に限定し，着目地域をつくば市天王台（〒 3001253）から距離が 0.05 の地域に指定する．これにより，空間ノードが 179 個，ウェブページと空間ノードの間の空間リンクが 329 個，空間ノードの間の空間リンクが 822 個生成された．

表 3: 提案手法によるハブ度の上位６件（豊島区）

doc_id	URL	サイト名
13039139	www.tabi-canada.co.jp/etour/Assets/commhtm/agent/tokyo-toshima.html	旅行代理店リスト　東京都豊島区
13513770	www.tokyo-kanteishi.or.jp/komatta/03_116.html	社団法人　東京都不動産鑑定士協会
7160587	www.hypermedia.or.jp/clp/kojin/iryo/zaikaku_kaigo/tosima.html	在宅介護支援センター　豊島区
6641959	www.h-nakano.ne.jp/about/teikei.html	提携先医療機関
10565146	www.nichi-bei.co.jp/shop/Hanbaiten_html/TokyoToshima/TokyoToshima.html	商品取扱店～豊島区エリア
10585207	www.nicos.co.jp/tokyo/contents/mado2.html	NICOS ギフトカード販売窓口

表 4: 提案手法によるハブ度の上位６件のリンク情報（豊島区）

doc_id	spa_link	effec_spa_link	weblink	effec_weblink	out_ratio	evaluation
13039139	8	6	0	0	0.7777	○
13513770	11	6	0	0	0.5833	○
7160587	12	5	0	0	0.4615	△
6641959	8	2	8	8	0.6470	×
10565146	8	2	0	0	0.3333	×
10585207	20	3	0	0	0.1904	×

提案手法によるハブ度の上位 6 件の実験結果を表 7 に示す．各ページのリンク情報を表 8 に示す．ハブ度が 3 位になるページ 9795990 はつくば市の国立研究機関，公益法人，民間研究機関，学校などのホームページのリンク集である．このページには空間情報がないが，有効なウェブリンクを有している（リンク先のページは指定された地域内の空間情報を有している）ため，良いハブページと見なされる．

空間ノード間に空間リンクを張るかどうかの閾値の影響を比較するために，同じく着目地域をつくば市天王台（〒 3001253）から距離が 0.05 の地域に指定し，空間ノード間に空間リンクを張るかどうかの閾値を 0.007 に増やした．これにより、同じく空間ノードが 179 個，ウェブページと空間ノードの間の空間リンクが 329 個生成された．閾値が増えたため，空間ノードの間の空間リンクは 1086 個に増えた．

提案手法によるハブ度の上位 6 件の実験結果を表 9 に示す．各ページのリンク情報を表 10 に示す．空間ノードに空間リンクを張るかどうかの距離の閾値を増やすことで，ウェブリンクを持たないが空間リンクを多数有しているページがより上位となっている．ハブ度が 1 位になるページ 4968622 はつくば市公民館講座の案内のページである．このページは 16 個の郵便番号を含んでいるが，そのうちの一つには対応する経緯度が算出できなかった（データベースに登録がなかった）ため，表 10 で示されたようにページの空間リンクの数は 15 である．そのうち 12 本の空間リンクは指定された地域内への空間ノードを指している．なお，今回の実験では，同じサイト内のウェブリンクを削除したため，このページのウェブリンクの数は 0 である．このページは空間情報を多く持ち，指定された地域内の空間情報の割合も高いため，良いハブページと見なされる．

5　まとめと今後の課題

本研究では，特定の地理領域に関し有用なリンクを張ってい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を提案した．また，実験により，out_ratio，in_ratio の有用性や，空間リンク作成のための閾値設定の影響を見た．

以下，今後の課題を挙げる．

1. 提案手法の改良
   (a) in/out_ratio 定義式の工夫：今回の提案手法では，有効なウェブリンクと空間リンクの割合を一つの指標にまとめ計算したが，それらを分離してウェブリンクの in/out_ratio，空間リンクの in/out_ratio とすることも考えられる．
   (b) リンクの重み付け：今回の提案手法では，ウェブリンクと空間リンクを同様に扱ったが，ウェブリンクと空間リンクに異なる重みをつけることも考えられる．
   (c) 閾値の決定：今回の実験では，空間ノード間に空間リンクを張るかどうかの閾値は適当に設定したが，適切な値を自動的に設定する方式の開発が求められる．
2. HITS 以外のアルゴリズム：今回の提案手法では，HITS をベースとしたが，HITS 以外のアルゴリズムを拡張することも検討したい．
3. データの収集：今回の実験では，NW100G-01 データを利用したが，着目する地域について，

表 5: *in/out_ratio* を導入しない手法によるハブ度の上位６件（豊島区）

doc_id	URL	サイト名
10822643	www.noris.co.jp/diving/tokyo/tokyo23-1.htm	東京 23 区のダイビングショップ&サービス
9321218	www.library.metro.tokyo.jp/16/16a00.html	都内公立図書館一覧
11724304	www.recycle.gr.jp/link.html	リサイクルソリューション
13039139	www.tabi-canada.co.jp/etour/Assets/commhtm/agent/tokyo-toshima.html	旅行代理店リスト　東京都豊島区
13513770	www.tokyo-kanteishi.or.jp/komatta/03_116.html	社団法人　東京都不動産鑑定士協会
7160587	www.hypermedia.or.jp/clp/kojin/iryo/zaikaku_kaigo/tosima.html	在宅介護支援センター　豊島区

表 6: *in/out_ratio* を導入しない手法によるハブ度の上位６件のリンク情報（豊島区）

doc_id	spa_link	effec_spa_link	weblink	effec_weblink	out_ratio	evaluation
10822643	62	4	26	26	0.3483	×
9321218	371	3	22	22	0.0659	×
11724304	209	4	31	31	0.1493	×
13039139	8	6	0	0	0.7777	○
13513770	11	6	0	0	0.5833	○
7160587	12	5	0	0	0.4615	△

このデータの中に必ずしも良い空間ハブがないこともあった．また，リンク先のページが必ずしもデータセットに含まれていないため，精度の低い解析となることも考えられる．データを有効に収集するための特定の領域に関する空間情報のクローリングも今後の課題である．

謝辞

本研究の一部は，日本学術振興会科学研究費基盤研究 (C)(2)(16500048)，基盤研究 (B)(15300027)，および CREST「自律連合型基盤システムの構築」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).

[12] 張建偉, 石川佳治, 北川博之, 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, 電子情報通信学会総合大会 (2004).

表 7: 閾値が 0.005 であるハブ度の上位 6 件（つくば）

doc_id	URL	サイト名
6608658	www.gsi.go.jp/LINK/GOVERNMENT/index.htm	各省庁のホームページへリンク
6191997	www.forum.esto.or.jp/open/link/link.html	
9795990	www.mexttci.go.jp/tsukuba_sites_j.html	つくばのサイト・リスト
7511229	www.inpaku.pref.ibaraki.jp/link/outline.html	いばらきバビリオン「未来科学館」
13941808	www.tsukubasyoko.or.jp/town.html	つくば市筑波商工会
8254577	www.jsokuryou.jp/list.asp	一覧 LIST

表 8: 閾値が 0.005 であるハブ度の上位 6 件のリンク情報（つくば）

doc_id	spa_link	effec_spa_link	weblink	effec_weblink	out_ratio	evaluation
6608658	0	0	630	45	0.0729	×
6191997	0	0	154	24	0.1612	×
9795990	0	0	92	17	0.1935	○
7511229	0	0	30	9	0.3225	△
13941808	1	1	10	10	1	○
8254577	0	0	43	1	0.0454	×

表 9: 閾値が 0.007 であるハブ度の上位 6 件（つくば）

4968622	www.city.tsukuba.ibaraki.jp/inform/educate/kominkan/list.htm	つくば市公民館講座のご案内
12469677	www.shokabo.co.jp/keyword/old/labo2000.html	今月のキーワード（研究所の一般公開）
10659761	www.nilim.go.jp	国土交通省国土技術政策総合研究所
5604499	www.e-golf.co.jp/tsukubane-cc/	つくばねカントリークラブ
8261844	www.jsurvey.jp/link.htm	測量関係リンク
4968829	www.city.tsukuba.ibaraki.jp/life/regist/section.htm	各庁舎市民窓口課一覧

表 10: 閾値が 0.007 であるハブ度の上位 6 件のリンク情報（つくば）

doc_id	spa_link	effec_spa_link	weblink	effec_weblink	out_ratio	evaluation
4968622	15	12	0	0	0.8125	○
12469677	52	10	0	0	0.2075	×
10659761	1	1	83	6	0.0941	△
5604499	1	1	4	1	0.5	△
8261844	33	4	1	1	0.1714	×
4968829	5	3	0	0	0.6666	○