地域ウェブ情報源の収集のためのクローリング手法の提案

張　建偉[†]　石川　佳治[†,††]　黒川　沙弓[†††]　北川　博之[†,††]

† 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
†† 筑波大学計算科学研究センター
††† 筑波大学第三学群情報学類
〒 305-8573 茨城県つくば市天王台 1-1-1
E-mail: †{zjw,saku39}@kde.cs.tsukuba.ac.jp, ††{ishikawa,kitagawa}@cs.tsukuba.ac.jp

あらまし　本稿では，ある特定の地域に関する情報をウェブから収集するクローリング手法の提案を行う．地域情報のクローリングにおいては，地域情報として有用なページが必ずしもウェブ上で人気が高いとは限らないこと，また，あるページが指定された地域に関連するか否かを機械的に判別することが難しいなどの問題点がある．本稿で提案する地域情報のクローリングの枠組みでは，対象の地域に対してユーザがすでに保持する実データの活用を図り，ウェブとデータベースの連携により探索処理を進める．提案手法では，まず，データベース中のデータを用いて，対象の地域に特化した小規模なグラフ構造をまず構築する．これにリンク解析を適用し，その結果をクローリング時に探索すべき次のページの選択に利用する．初期実験において,「つくば市の飲食店」を元データとしたクローリングに対して，次に探索対象となる候補ページの上位につくば市のポータルサイト，およびグルメ関連のサイトが来ることが確認された.

キーワード　クローリング，クローラ，地域情報，リンク解析，ウェブとデータベースの統合

A Web Crawling Method for Collecting Local Web Sources

Jianwei ZHANG[†], Yoshiharu ISHIKAWA[†,††], Sayumi KUROKAWA[†††], and Hiroyuki KITAGAWA[†,††]

† Department of Computer Science, Graduate School of Systems and Information Engineering,
†† Center for Computational Sciences,
††† College of Information Sciences, Third Cluster of Colleges,
University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki, Japan, 305-8573
E-mail: †{zjw,saku39}@kde.cs.tsukuba.ac.jp, ††{ishikawa,kitagawa}@cs.tsukuba.ac.jp

Abstract　In this paper, we propose a web crawling method to collect web pages related to a specific geographical local area. There are some problems on crawling local web pages. First, a useful local page is not necessarily a popular page. Second, it is difficult to decide whether a page is related to the specified area or not. In our framework, we try to integrate the web and a user database containing information on a specified area. We first construct a small linked graph structure for the specified area using entries contained in the database. Next a link analysis is performed, then the result is used by the crawler to select the web page to be crawled next. With the help of the real data on the local area, the crawler performs an effective web crawling. As a preliminary experiment, we have simulated a web crawling for restaurants in Tsukuba city. It is shown that some portal sites for Tsukuba and pages related to gourmet restaurants are coming to the top place as the crawling candidates.

Key words　crawling, crawler, local information, link analysis, integration of web and databases

1. はじめに

ウェブ上の情報の活用において，近年着目されている話題の一つに**ローカルサーチ**（local search）がある．特定の地域に特化したウェブ検索機能を提供し，ある地域に興味を持つユーザへの適応的なサービスを提供しようというものである [1], [2]. この種のサービスを実現するためには，特定の地域に関するウェブページを大規模なウェブの中から効率よく，高い精度で抽出する機構が求められる．ウェブページを効果的に収集するクローリング手法に関しては，これまでさまざまなアプローチが提案されており，特に，ユーザが着目しているトピックに関するウェブページを集中的に収集する**トピック主導型クローリ**

ング（topic-focused crawling）については，数多くの研究がなされている [3]〜[5]．しかし，地域情報のクローリングに関しては，次のような問題がある．

（1） ページの地域性の判定の難しさ：地名にはあいまい性が存在するため，ページ中に出現する字句のみで地域を特定することは必ずしも容易ではない．たとえば，複数の土地に同じ地名が用いられる場合や，人名などの固有名詞と地名が一致する場合（例：千葉さんの個人ページ）や，地名が一般名詞として出現する場合（例：札幌ラーメン）もある．また，異なる地域に関する情報がページ内に複数現れるページ（例：支店一覧のページ）もある．

（2） ウェブ全体でのページの評価の有用性の問題：トピック主導型クローリングにおいては，トピックに関連し，かつ，評判の高いページを優先的に探索することが一般的な戦略である．しかし，ある地域に関して重要な情報を提供しているページであっても，被リンク数が少ないなどの理由から，ウェブ全体を考えた場合には評判が必ずしも高くはないという問題がある．

（3） 地域情報に関連した実データの有効利用：ある特定の地域を考えた場合，その地域内に存在する組織のウェブサイトや個人のウェブページのように，実世界に存在する実体と関連付けできるウェブ情報源も存在する．これらの情報源の一部については，ユーザがすでに知っていたり，ウェブディレクトリなどで容易に発見できる場合も多いと考えられる．そのため，その地域に関する既知の情報を有効に活用してウェブページの探索に利用する枠組みが重要であると考えられる．

そこで本研究では，ユーザがすでに保持している着目地域に関する実データを有効利用したウェブクローリング手法を提案する．ウェブページの地域性の判定結果やウェブページ間のリンク情報のみでなく，実データとの関連性も考慮して，探索するページの選択を行う．また，ユーザからのフィードバックによる実データの拡充などを行い，選択的なクローリングを実現する．

2. 関 連 研 究

2.1 ウェブのクローリングとリンク解析

ウェブのクローリング方式においては，従来よりさまざまな方式が工夫されてきた [3], [4]．効率的なクローリングには重要なページを優先的に辿ることが求められる．そのための手法の一つとして，PageRank クローラ [6] がある．これは，PageRank [7] によるスコアに基づいてページのアクセスの順序を決めるアプローチである．

クローリングの効率化に加え，近年，**トピック主導型クローリング**（topic-focused crawling）が着目されている [5], [8], [9]．これは，与えられたトピックに対するページを効率よく収集することを目的としている．[8], [9] では，分類器を用いて選択的にクローリングを行う手法が提案されており，[5] では複数のトピック主導型クローリングにおける能力の比較を行っている．

本研究の目的は，実データに関連するウェブページを優先的に効率よく探索し，実データとの融合を図る点にあり，トピック主導型クローリングと関連が深い．しかし，実データに関連するページはウェブ上では必ずしも評判が高いとは限らず，わずかな参照しか存在しない場合も多い．そのため，提案手法では，実データに密接に関連するページをサーチエンジンなどを用いて初期探索し，局所的なグラフに対してリンク解析を行い，これをもとにクローリング時のページ選択に用いる．

ウェブのリンク解析にヒントを得て，近年ではデータベースに対してもリンク解析を適用するアプローチが提案されている [10], [11]．データベース中に存在する関連情報を一種のリンクとみなしてリンク解析を行うもので，データベースの問合せ能力の活用などが課題となっている．本稿で提案する手法では，実際のデータベースを外部のウェブページと統合してリンク解析を行うという点で，これらのアプローチの拡張になっていると考えることができる．

2.2 地域情報の収集と検索

ウェブの中から特定の地域に関するページを抽出し検索などに利用するための研究が進められている．本研究に特に関連する研究をいくつか紹介する．[12] では，位置指向検索に必要なウェブページの選択的な収集を行い，位置指向のウェブページ検索を実現するシステムを開発した．リンク文字列に地域情報が含まれるかどうかにより，参照先のページの内容を予測し，地域情報を優先的に収集する手法を提案している．また, [13] では，ウェブページの内容の偏在性，話題の遍在性とユーザの偏在性を考慮して，ページのローカル度の抽出手法を提案している．ローカル度はウェブページがどの程度地域に密着しているかの判断に利用する．

[14] 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．本論文ではリンク解析をクローリングに利用する点や，ウェブ情報と実データを統合利用する点が異なる．一方, [15] 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手法を述べている．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．実データとウェブ空間の関連付けという点では本研究と共通性がある．一方，我々は [16] において，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って拡張し，拡張されたグラフ構造におけるリンク解析により，空間情報ハブを抽出する手法を提案している．

3. 提 案 手 法

3.1 基本的なアイデア

トピック主導のクローリングにおいては，シードとして与えられたページ群をもとに，リンクによる参照とページの内容に基づいてページが選択的に収集される．一方，提案手法におけるクローリングは，ユーザにより提供される，ユーザが興味を持っている対象地域の実データをもとに行う点に特色がある．この意味で，本アプローチはデータベース主導のクローリングのアプローチと位置づけられる．また，従来のクローリングでは，できるだけ多くのページを収集することが基本であったのに対し，提案手法では，ユーザが提供するデータベースをウェ

ブ上の情報をもとに補完・拡充することに焦点を当てている点が異なる．

3.2 データベースの拡張

具体例として，あるユーザがつくば市内のレストラン情報に興味を持っているとし，図 1 に示すようなレストラン情報を保持しているものとする．

class	name	URL	address	phone	zip-code
フレンチ	ビストロ A	foo.jp/~xxx/	つくば市○○	···	···
すし	B 寿司	bar.jp/~yyy/	つくば市××	···	···
⋮	⋮	⋮	⋮	⋮	⋮

図 1　テーブルの例：`restaurant`

提案手法では，このようなテーブルからなるデータベースに対し，クローリング処理を設定する中間的な段階として，まず，図 2 に示すような**拡張データベース**（extended database）を考える．これは，実データとウェブページの関係を実体関連モデル風に表現したものであり，左側が実データベースを表し，右側がウェブを表す．実体集合 `restaurant` と `page` は関連集合 `has-HP` で関係付けられる．なお，`has-HP` 関連集合に付与された“ (1, 1) ”および“ (0, 1) ”という表記は，各レストランにはただ 1 つのページが必ず対応するが，すべてのウェブページがレストランのホームページに対応するわけではないことを示している．実体集合 `page` はウェブページの全体集合を表し，各ページ自身の URL とそのページ中に含まれる住所表記，郵便番号，電話番号を属性としてとり，リンクによる参照を表す多対多関連 `refers` で自身に関連付けられる．なお，2 重の矢印は，その属性が 0 個以上複数の値を持つ多値属性であることを略記している（この意味で実体関連モデルの表記を若干拡張している）．

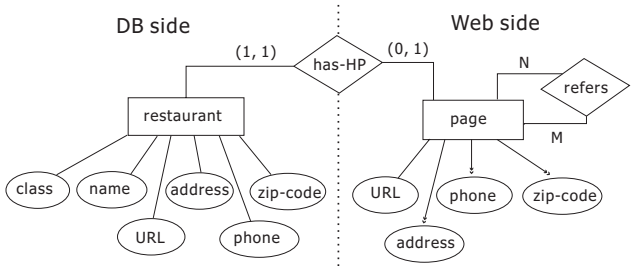

図 2　拡張データベース

他のデータに対して拡張データベースを構築する際も，同様のモデル化を行うが，この例のようにウェブページの属性として必ず電話番号や郵便番号などの特徴を抽出しなければいけないというわけではない．ユーザの要求および利用可能なツールなどに合わせて，ウェブページから抽出する属性は変更・拡張が可能であるものと考える．

3.3 オーソリティ伝播グラフ

次いで図 3 に示すような**オーソリティ伝播グラフ**（authority transfer graph）を作成する．このグラフは図 2 に示した拡張データベースの構造をもとにグラフの要素間の関連を表すもので，クローリングのための方針を指定するために作成する．なお，

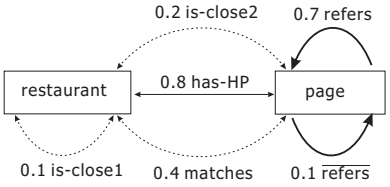

図 3　オーソリティ伝播グラフの例

属性については省略している．

実線・破線の矢印（**リンク**（link）と呼ぶ）は，ウェブページ間のリンクの概念をデータベースにまで拡張したものに相当する．実線のリンクは，図 2 中の関連集合に相当する．リンクに付与された数値（0 から 1 の間の値）をリンクの**重み**（weight）と呼ぶ．たとえば，`refers` というリンクは，あるページから他のページへの URL による参照に相当し，付与されている 0.7 という数値は，参照元のページの重みに 0.7 を掛けた値が参照先のページに渡ることを意味している．これは PageRank [7] などに見られるリンク解析のアイデアに基づいている．$\overline{\texttt{refers}}$ のように上線がついたリンクは，逆方向のリンクを表し，この例の場合は，参照されているページの重みに 0.1 を掛けた値が参照元に渡ることを意味する．両方向への重みが同じ場合には，`has-HP` のように ⟷ で表現する．

破線のリンクは仮想的に構築される関連を表す．`restaurant` 自身をリンクする `is-close1` は，2 つのレストランが近くに位置する場合にのみ発生する関連である．一方，`restaurant` と `page` をリンクする `is-close2` は，レストランの住所とページが表す地域が近い場合にのみ発生する関連である．`matches` は，レストランの情報とウェブページ内の情報が関連しているとコンテンツ解析などをもとに判定されたときのみに発生する関連である．

仮想的な関連は，一般に曖昧性を含んだ関係を表しており，具体的な実装レベルでその定義が与えられると考える．たとえば，`is-close1` については，`restaurant` テーブルの `address`, `phone`, `zip-code` 属性をもとに，位置が近いことを判定するユーザ定義述語をデータベースに導入することが考えられる．具体的な実装のアイデアについては後述する．

上記の `is-close1` などの関連については，関連があるか否かという 2 値的な判断をしているが，たとえば距離に応じてリンクの重み付けに反映するなどの拡張も考えられる．これについては今後の課題としたい．

3.4 リンク解析

本手法では，ウェブのクローリングを開始する前に，オーソリティ伝播グラフをもとに実データとそれに関連するウェブページのリンク構造を解析する．リンク解析の結果を用いることで，ウェブ上の関連情報のクローリングの効果を向上させることを目的とする．前節の例を用いて説明する．

3.4.1 グラフ構造の構築

まず，グラフの各ノードを生成する．

（1） `restaurant` テーブルの各行に対しノードを生成する．

（2） restaurant テーブルの各行に対し URL 属性の値をもとにホームページを取得する．得られたホームページの集合を S_1 とする．

（3） 各ホームページから辿れる同一サイト内（ここで同一サイト内とは，ホームページの URL をプレフィックスに持つような URL のページ）のページも収集し，集合 S_2 とする．

（4） 各ホームページに対し，サーチエンジンを用いてバックリンクを調べ，ページ集合 S_3 を取得する．

（5） restaurant テーブルの各行の値をもとに，外部のサーチエンジンへ問合せを発行し，結果のページを取得する．たとえば，name 属性と phone 属性の値をサーチエンジンに検索条件として与え，結果のページ群をダウンロードし，集合 S_4 とする．

（6） ステップ 2, 3, 4, 5 で得られた集合 S_1, S_2, S_3, S_4 の各ページについて，ノードを生成する．

（7） ページ集合 $S_1 \cup S_2 \cup S_3 \cup S_4$ の各ページについて，リンクの抽出を行い，未ダウンロードの URL に対しても，ノードを生成する．

以上のように，グラフのノードは restaurant テーブルのエントリ，それらに関連するダウンロード済のページおよびページから抽出した未ダウンロードの URL からなる．

次に，各ノードの間のリンクを生成する．

（1） 集合 S_1 のページに対し，対応する restaurant ノードとの間に双方向の has-HP のリンクを生成する．

（2） ノード生成についてのステップ 7 のリンク抽出結果を用いて，ダウンロードしたページ間，ダウンロードしたページと未ダウンロードの URL の間に，refers と $\overline{\texttt{refers}}$ のリンクを生成する．

（3） S_4 の各ページに対し，対応する restaurant ノードと双方向リンク（matches）で接続する．

（4） restaurant テーブル上でユーザ定義述語 is-close1 を用いた self-join を評価する．評価結果の各ペアについて，テーブルのエントリ間に双方向リンク is-close1 を作成する．ユーザ定義述語 is-close1 は，たとえば，属性 address の値を座標値に変換した後，距離を評価し，その値が閾値（例：1km）以下の場合に真となるように実装する．

（5） ページ集合 $S_1 \cup S_2 \cup S_3 \cup S_4$ の要素のページについて，restaurant テーブルの各行との間に is-close2 の条件が成り立つかをチェックし，成り立つ場合は双方向リンクを作成する．is-close2 の判定は，たとえばページ内に含まれる住所，郵便番号などを用いて行う．

ノード間の辺については，オーソリティ伝播グラフでの指示に基づく重みを付与する．なお，曖昧さが生じる場合のみ辺にラベルを付与する．重みの設定においては，同一ノードから同じ関連により複数の辺が出る場合には，辺の数により重みを等分する．たとえば，図 4 では「ビストロ A」から 2 本の matches 辺が出ているため，それぞれの辺の重みは図 3 における重み 0.4 を等分して 0.2 となる．一方，参照先のページから「ビストロ A」への逆方向の辺の重みは 0.4 のままとなる．これは PageRank [7] の考え方に基づいている．

以上のようなアプローチにより，図 4 のようなデータベース中のデータに関連の深いページ群からなるグラフ構造を得る．このようなグラフを**データグラフ**（data graph）と呼ぶ．

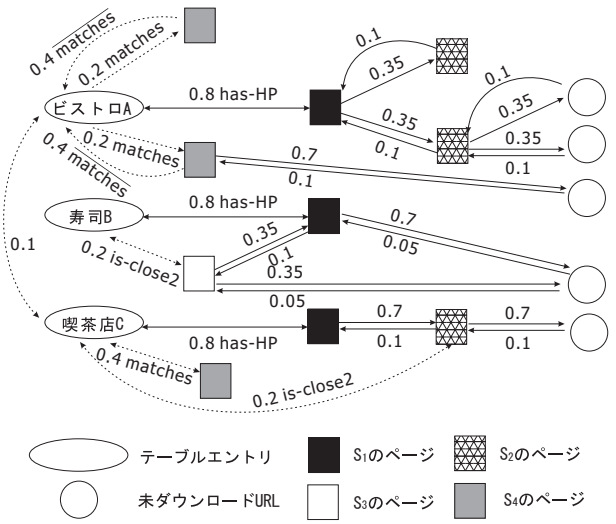

図 4　データグラフの例

3.4.2　スコアの計算

次いで，前節で構築されたデータグラフにおいて，各ノードの **LocalRank** 値を計算する．LocalRank 値はデータベースエントリあるいはページ (URL) と対象地域との関連度を表すスコアである．下記の式で計算する．

$$\boldsymbol{r} = d\mathbf{A}'\boldsymbol{r} + \frac{(1-d)}{n}\boldsymbol{e} \tag{1}$$

n はデータグラフのノード数である．$\mathbf{A}$ は，$n \times n$ 行列であり，その (i, j) 要素はノード i からノード j への辺の重みの総和である．$\mathbf{A}'$ は，$\mathbf{A}$ の転置行列を表す．$\boldsymbol{r}$ は n 次元列ベクトルであり，各ノードの LocalRank 値を表す．$\boldsymbol{e}$ はすべての要素が 1 である n 次元列ベクトルである．d はユーザから指定される定数（*damping factor*）であり，辺に沿ったノード間の移動とランダムな移動を調整する．LocalRank 値は，$\boldsymbol{r}$ の各要素を 1 に初期化し，上の式を収束するまで評価することで得ることができる．LocalRank の計算式は PageRank と同様だが，計算は通常のウェブグラフではなく，上記のデータグラフにおいて行うという点が異なる．

3.5　クローリング処理

クローリング処理においては，前節でその導出法を述べた LocalRank によるスコアを利用する．アルゴリズムの概要をアルゴリズム 1 に示す．このアプローチは，LocalRank 値により次にアクセスするページを決めることから，いわゆる PageRank クローラ [5], [6] の拡張となっている．

未アクセス URL の集合の中から，LocalRank 値が最大の URL を選んでアクセスした後，アクセスしたページと各データベースノードとの matches，is-close2 のリンクおよびウェブノードとの refers のリンクを追加して LocalRank 値を再計算し，次のアクセス先のページを決めるという繰り返しが基本となる．ただし，matches のリンクの追加（10〜12 行目）は，コンテンツ解析によりページの内容がデータベースノードに

マッチしたと判断される場合，ページとデータベースノード間の matches のリンクを追加する．is-close2 のリンクの追加（13〜15 行目）は，ページ内の空間情報を抽出し，データベースノードとの距離を計算して，近い（ある閾値以下）と判断される場合，is-close2 のリンクを追加する．これにより，仮想的なリンクの導入も動的に行っている．なお，省略しているが，それぞれ逆方向のリンクの追加も行われる．

Algorithm 1 LocalRank Crawling

```
1: 前節の手法により，各ページ（URL）の LocalRank 値を計算する．
2: アクセスした URL の集合を crawled_urls とする．
3: 未アクセスの URL の集合を target_urls とする．
4: while not empty(target_urls) do
5:    target_urls 中で LocalRank 値が最大の URL を url とする．
6:    url を target_urls から削除する．
7:    page := get_page(url)
8:    append(crawled_urls, url)
9:    for all d ∈ DB do
10:      if matches(d, page) then
11:         append(matches, (d, url))
12:      end if
13:      if is-close2(d, page) then
14:         append(is-close2, (d, url))
15:      end if
16:   end for
17:   url_list := extract_urls(page)
18:   for all u ∈ url_list do
19:      append(refers, (url, u))
20:      if u ∉ crawled_urls and u ∉ target_urls then
21:         target_urls := target_urls ∪ {u}
22:      end if
23:   end for
24:   各ページ（URL）の LocalRank 値を再計算する．
25: end while
```

4. 実　　験

4.1 データベースの内容と初期ページの収集

この実験では，つくば市の飲食店に焦点を当てたクローリングを考える．まず，実データを表すデータベースとして，ウェブからつくば市内に存在し，ホームページを有する 54 件の飲食店を人手で抽出した．ホームページから住所，電話番号，郵便番号が入手できる場合には，それらの情報も入手した．図 5 にその例を示す．

データベース中の各データについて，関連するページを次のような方針で収集した．

（1） ホームページ，および，ホームページから辿ることができる同一ドメインのページを収集する：結果として，1,463 個のページが取得された．

（2） ホームページを参照しているページの URL （バックリンク）をすべて収集する：収集には Google を利用し，166 個のバックリンクを収集し，ページをダウンロードした．

（3） 飲食店の名称と電話番号をキーワードとしてサーチエンジンに与え，関連するページ群を収集する．具体的には，Google に“ コッコリーノ 029-864-4555 ”などの検索条件を与え，結果の 1 ページ目の URL をすべて抽出した．285 個の URL が得られ，これらをダウンロードした．

1 から 3 のページ数を加算すると 1,894 であるが，URL に重複があるため，ダウンロードしたページの総数は 1,812 であった．

4.2 ページからの情報抽出とグラフの構築

ダウンロードした各ページについて，以下のような方針で情報の抽出を行った．

（1） リンクの抽出：タグで表される他ページへの参照を抽出する．同一ページへの参照やメディアファイルへのリンク，CGI ページへのリンクなどは抽出の対象から外す．

（2） 郵便番号の抽出：“ 〒 305-8573 ”のように，“ 〒 ”に 7 桁の数字が続く文字列（ハイフンは省略可）をそのページに対する郵便番号として抽出する．ただし，ページ内に複数個の郵便番号が出現する場合には，あいまい性が生じるため採用しない．つまり，ページ内に郵便番号がただ 1 つだけ含まれる場合にのみ抽出を行う．

（3） 住所表記の抽出：ページ内から“ 茨城県つくば市天王台 1-1-1 ”という形式で表される住所表記を抽出する．ただし，正確さのため，都道府県名からはじまり番地レベルまで指定してあるフル表記の住所のみを対象とする．なお,「一丁目」などの，漢字による番地表記については，パターン照合の困難さもあり対応からは省く．この場合にも，ページ内に住所表記がただ 1 つだけ含まれる場合のみ，抽出を行う．

（4） 上記 3 で抽出された住所表記に対し，その緯度と経度を求める：“ Yahoo!地図情報 ”[17] を用いて，緯度経度の情報を検索する．ただし，このサイトではデータの不足などにより，一部の住所については座標値を返さない．このような住所に関しては，今回は手作業でもっとも近い住所の座標値を選択する．なお，座標値の算出は，図 5 の各データについても行う．

以上の結果をもとに，解析に用いるノードとリンクの情報を以下のように抽出した．

（1） ノードとしては，飲食店オブジェクトに対応するノードと，各 URL （ダウンロード済のページと未ダウンロードのページの両者を含む）に対するノードを作成する．前者については 54 個，後者については 12,952 個を生成し，合計で 13,006 個のノードが得られた．

（2） has-HP 関連に相当するリンクについては，飲食店とそのホームページという関係から，54 個のリンクを生成する．

（3） ページ間の参照に関する refers 関連については，上記のリンク抽出結果を利用する．ただし，同一ドメイン内のリンクについては対象外とする．この結果，8,663 個のリンクが得られた．

（4） matches のリンクについては，飲食店の名称と電話番号を用いて，Google から取得したページと飲食店オブジェクトの間の内容関連を表し，285 個のリンクが作成された．

（5） 飲食店オブジェクト間の近接関連を表す is-close1 については，

(a) オブジェクトの座標をもとに計算した距離が 2km 以内

restaurant

ID	名前	URL	住所	電話番号	郵便番号
1	コッコリーノ	http://cocclino.jp/	茨城県つくば市筑穂 3-1-5	029-864-4555	300-3257
2	グルマン	http://www.omisemall.com/goru/	茨城県つくば市吾妻 3-7-17	029-851-6107	305-0031
⋮	⋮	⋮	⋮	⋮	⋮

図 5　飲食店データベース

である場合，もしくは

(b) オブジェクトの郵便番号が同じ場合

にリンクを作成するものとする．結果として 453 個のリンクが作成された．なお，緯度経度に基づく距離の計算には，国土地理院提供がウェブ上に提供している計算システム [18] を利用した．

（6） オブジェクトとページの間の近接関連を表す `is-close2` についても同様に，

(a) オブジェクトの座標とページの座標（住所表記が抽出できた場合）をもとに計算した距離が 2km 以内である場合，もしくは

(b) オブジェクトの郵便番号とページの郵便番号が同じ場合

にリンクを作成するものとする．結果として 1,407 個のリンクが作成された．

4.3 リンク解析

LocalRank 法に基づくリンク解析は，3.4.2 節の式 (1) で十分に収束するまで計算を行った．先述のように，ノードの総数は $n = 13,006$ である．d は $d = 0.9$ と設定した．$\mathbf{A}'$ は $\mathbf{A}$ の転置行列である．行列 $\mathbf{A}$ の (i, j) 要素 a_{ij} は，ノード i からノード j に対しての重みを表す．$\mathbf{A}$ の値を以下のように設定する．

（1） ノード i, j が `has-HP` で関連しているとき，a_{ij} に 0.8 を加える．なお，a_{ji} にも 0.8 を加える．

（2） ノード i が `refers` 関連で他の m 個のノード $n_1, n_2, \ldots, n_m$ を参照しているとき，a_{i,n_j} $(j = 1, \ldots, m)$ に $0.7/m$ を加える．一方，ノード j が `refers` 関連で他の m' 個のノード $n_1, n_2, \ldots, n_{m'}$ から参照されているとき，a_{j,n_i} $(i = 1, \ldots, m')$ に $0.1/m'$ を加える．

（3） ノード i が `matches` 関連で他の m 個のノード $n_1, n_2, \ldots, n_m$ と関連しているとき，a_{i,n_j} $(j = 1, \ldots, m)$ に $0.4/m$ を加える．

（4） ノード i が `is-close1` 関連で他の m 個のノード $n_1, n_2, \ldots, n_m$ と関連しているとき，a_{i,n_j} $(j = 1, \ldots, m)$ に $0.1/m$ を加える．

（5） ノード i が `is-close2` 関連で他の m 個のノード $n_1, n_2, \ldots, n_m$ と関連しているとき，a_{i,n_j} $(j = 1, \ldots, m)$ に $0.2/m$ を加える．

リンク解析処理の計算は，疎行列操作機能を有する行列計算ソフトウェア Scilab [19] を用いて行った．実際の計算時間は，3 秒程度であった．ただし，データを読み込むのに別途数十秒程度の時間を要した．

4.4 解析結果

まず，飲食店オブジェクトをランク値順に並べたものの上位 8 件と下位 9 件を図 6 に示す．支店が存在しないようなローカルな飲食店が比較的に上位となっている．すき家，吉野家といった全国のチェーン店のお店はランクが低くなっている．ローカルな飲食店ほど，ホームページに近いところに住所や電話番号が書かれているのでランクが高くなっているのではないかと考えられる．これは，地域情報が上位に来る結果となっているのでいい結果と言える．

すでにダウンロード済のページについて，ランク順に上位 20 件を取り出したものが図 7 である．確認してみると，上位のページは飲食店オブジェクトの一位の「蘭亭」に関連するページ，あるいは,「蘭亭」に言及しているページが多く見られた．たとえば，上位 20 件では，1-5, 8-15, 18 番目のページがこれにあたる．他の上位のページは，他の上位ランクの飲食店の関連ページなどが主であった．また，上位 50 件のページを調べると，そのうちつくば市の飲食店情報に関連していないページはわずか数件程度であり，つくば市のローカルなレストラン情報が集まっていると言える．

未ダウンロード URL について，ランクの上位 20 件を取り出したものが図 8 である．なお，長い URL については末尾を省略している．これらのページを説明すると，以下のようになる．

- 1 件目：Apos つくば（つくば市の地域情報ポータルサイト）
- 2 件目：ぐるなび：レストラン・飲食店検索
- 3 件目：ぐるなび関東版
- 4 件目：ぐるなびの格付けページ（パスワード付きアクセス）
- 5 件目：国土交通省国土技術政策総合研究所
- 6 件目：アイキューブつくば（つくばの人材派遣）
- 7 件目：Google の検索ページ
- 8 件目：つくばマルチメディア社のホームページ
- 9 件目：1 件目と同一のサイト
- 10 件目：茨城県総合案内板
- 11 件目：つくば@TOWN（つくば市の地域バーチャルタウン）
- 12〜20 件目：ぐるナビ提供の店舗の地図

上位にはつくば市のポータルサイトやグルメサイトが来ており，つくば地域のローカルな情報をよく表しているといえる．ただし，調べると 1 件目，9 件目，10 件目，11 件目のサイトは 8 件目の企業により運営されており，相互にスコアを強めあう関係にあると考えられる．

実験結果を見る限り，一部を除き，上位につくば市を代表するローカルなサイトが位置づけられたことにより，クローリングの際のリンク対象の選択にとっては有効であったと考えられる．

今回の実験では，クローリング自体の処理は行わず，データベースの情報をもとに収集したページ群より LocalRank の計算を行い，次に探索対象となるページがどのようになるかを見た．その結果，つくば市のポータルサイト，およびグルメ関連のサイトが上位に来ることが確認された．これは,「つくば市の飲食店」を元データとしたクローリングに対しては妥当な結果ということができる．

5. 今後の課題と議論

5.1 まとめと今後の課題

本稿では，実データとウェブデータを統合的に用いて，地域的な情報を探索するクローリングのアプローチを示した．また，以下のような点が今後の課題として挙げられる．

- 大規模なデータを用いた実験：今回の実験では，54 個という限られた数のデータを対象としていたが，より多くのデータを用いた場合にどのような変化が現れるかを検証する必要がある．また，大規模な実験に耐えうるシステム実装技術の開発を図る必要がある．
- 今回の対象データは飲食店であったが，他のカテゴリ（例：学校，ショッピング，研究開発）などではまた異なる結果が得られる可能性がある．これらについても検証が必要である．
- ページ内容の抽出については，今回は同一ページ内に複数の郵便番号，住所表記が現れる場合には単一の郵便番号，住所表記のみの場合だけ採用するというアプローチをとった．複数の住所情報が出現する際の対応についても検討の余地がある．また，ページ内容のより詳細な解析も必要である．
- LocalRank 手法については，リンクに関する重みの違いによる影響の調査，および，適切な重みの選択手法などを開発することが求められる．
- 実データの拡充：クローリングを行う際に，集積されるウェブページの中には，データベースの拡充に利用できる有益なページも含まれていると考えられる．そのページの内容を既存のデータベースに追加し，その情報を反映することが考えられる．これ自体は自動化が難しく人手を介することが避けられないが，有用なページを判定する作業自体は LocalRank 値が参考になると考えられる．

5.2 エントリ固有の LocalRank の利用

提案手法は，実データとウェブデータの融合という観点から捉えることができる．この考え方を進めると，単に全体的なスコア値を計算するのではなく，データベース中のある特定のエントリに焦点を当ててスコア付けをすることも考えられる．このようなアプローチは，ObjectRank の論文 [10] では *keyword-specific ObjectRank* と呼ばれており，データベース中のある特定のオブジェクト（キーワード）に着目してスコア計算を行うものである．その考え方を本提案手法のコンテクストに当てはめると，データベース中のあるエントリ（例：着目しているレストラン A）に関する LocalRank は

$$\boldsymbol{r}_T = d\mathbf{A}'\boldsymbol{r}_T + (1-d)\boldsymbol{s} \tag{2}$$

と表現できる．ただし，$\boldsymbol{s}$ は，着目するエントリ（レストラン A）に該当するノードについてのみ 1 を設定し，残りを 0 とした n 次元列ベクトルである. [10] では，このようにして得られたエントリ固有の LocalRank と，式 (1) で定義される大域的な LocalRank を用いることで，ノード v のスコアを

$$r(v) = r_T(v) \cdot (r_G(v))^g \tag{3}$$

と計算するアイデアが示されている．ただし，$r_T(v)$, $r_G(v)$ はそれぞれ，ノード v に対するエントリ固有および大域的な LocalRank 値であり，g はユーザにより指定される定数である．このようなアプローチにより，特定のエントリに重点をおいたランク付けが可能となり，これは一種の問合せとして機能する．

謝　　辞

本研究の一部は，日本学術振興会科学研究費 (C)(2)(16500048), 旭硝子財団研究助成，稲森財団研究助成，及び CREST「自律連合型基盤システムの構築」による．

文　　献

[1] goo タウンページ. http://townpage.goo.ne.jp/
[2] Yahoo!地域情報. http://local.yahoo.co.jp/
[3] P. Baldi, P. Frasconi, and P. Smyth, *Modeling the Internet and the Web: Probabilistic Methods and Algorithms*, Wiley, 2003.
[4] S. Chakrabarti, *Mining the Web: Analysis of Hypertext and Semi Structured Data*, Morgan Kaufmann, 2002.
[5] F. Menczer, G. Pant, and P. Srinivasan, Topical Web Crawlers: Evaluating Adaptive Algorithms. *ACM Trans. on Internet Technology*, Vol. 4, No. 4, pp. 378-419, 2004.
[6] J. Cho, H. Garcia-Molina, and L. Page, Efficient Crawling through URL Ordering. *Computer Networks*, Vol. 30, No. 1-7, pp. 161-172, 1998.
[7] S. Brin and L. Page, The Anatomy of a Large-scale Hypertexual Web Search Engine. *Computer Networks*, Vol. 30, No. 1-7, pp. 107-117, 1998.
[8] S. Chakrabarti, M. van den Berg, and B. Dom, Focused Crawling: A New Approach to Topic-specific Web Resource Discovery. *Computer Networks*, Vol. 31, No. 11-16, pp. 1623-1640, 1999.
[9] M. Diligenti, F. M. Coetzee, S. Lawrence, C.L. Giles, and M. Gori, Focused Crawling Using Context Graphs. *Proc. VLDB 2000*, pp. 527-534, 2000.
[10] A. Balmin, V. Hristidis, and Y. Papakonstantinou, ObjectRank: Authority-Based Keyword Search in Databases. *Proc. VLDB 2004*, pp. 564-575, 2004.
[11] F. Geerts, H. Mannila, and E. Terzi, Relational Link-Based Ranking. *Proc. VLDB 2004*, pp. 552-563, 2004.
[12] 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.
[13] 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.
[14] 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.
[15] 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.
[16] 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. DBSJ Letters, Vol. 3, No. 3, pp. 9-12, 2004.
[17] Yahoo!地図情報. http://map.yahoo.co.jp/
[18] 測量計算（距離と方位角の計算）. http://vldb.gsi.go.jp/sokuchi/surveycalc/bl2stf.html
[19] Scilab Home Page. http://scilabsoft.inria.fr/

1	0.351643	蘭亭	http://e-tsukuba.jp/rantei/index.htm
2	0.002898	ホワイト餃子店 つくば支店	http://www.white-gyouza.co.jp/detail/detail18.htm
3	0.001897	AtoZ	http://www8.ocn.ne.jp/ atoz/
4	0.001506	La Carafe	http://carafe.midi.co.jp/
5	0.001068	グルマン	http://www.omisemall.com/goru/
6	0.000842	D-Pocket つくば店	http://dpocket.hp.infoseek.co.jp/
7	0.000831	源	http://www.genchan.jp/
8	0.000754	ほっと BB ステーションつくば学園西店	http://hotstation.ne.jp/shop-list/tsukuba.html
⋮	⋮	⋮	⋮
46	0.000053	すき家	http://www.zensho.com/
47	0.000053	吉野家	http://www.yoshinoya-dc.com/
48	0.000051	とん亭	http://www.geocities.co.jp/Foodpia-Olive/4171/
49	0.000050	香辛飯屋	http://www.5488.net/2002/top.cgi
50	0.000050	バーミヤン	http://www.skylark.co.jp/
51	0.000050	一太郎	http://www.ichitarou.com/
52	0.000050	H2O@CAFE	http://wing.zero.ad.jp/H2OCAFE/
53	0.000050	サイゼリヤ	http://www.saizeriya.co.jp/
54	0.000050	カプリチョーザ	http://www.capricciosa.com/

図 6　オブジェクトに対するランク計算結果

1	0.260874	e-tsukuba.jp/rantei/index.htm
2	0.021487	r.gnavi.co.jp/a275100/
3	0.018932	e-tsukuba.jp/rantei/link.htm
4	0.014155	www.geocities.jp/papy0164/syokuji/you.html
5	0.012105	r.gnavi.co.jp/a275100/map1.htm
6	0.012086	www.collaborate-ibaraki.jp/dirsearch/kigyou/namediv/ni.asp
7	0.012056	www.joyoliving.co.jp/kurashi/data/thisweek.php?category=recruit
8	0.011492	www.joyo-net.com/mise/mise040408.html
9	0.011492	www.capital-group.co.jp/rantei.htm
10	0.011492	www.piazza.ne.jp/piazza/gourmet/index.asp?mode=detail&id=116
11	0.007308	e-tsukuba.jp/rantei/recruit.htm
12	0.007285	e-tsukuba.jp/rantei/company.htm
13	0.007185	e-tsukuba.jp/rantei/osusume.htm
14	0.007185	e-tsukuba.jp/rantei/email.htm
15	0.007185	e-tsukuba.jp/rantei/traffic.htm
16	0.002886	www.iki-iki.net/v7/townpage/a-you.htm
17	0.002830	tarea.hp.infoseek.co.jp/lminami.html
18	0.002737	www006.upp.so-net.ne.jp/puni/ibaraki-r-w1.htm
19	0.002670	www.h3.dion.ne.jp/ b-gakuji/syaon.html
20	0.002266	www.white-gyouza.co.jp/detail/detail18.htm

図 7　ダウンロード済のページに対するランク計算結果（上位 20 件）

1	0.019240	www.i-tsukuba.com/index.shtml
2	0.002976	www.gnavi.co.jp
3	0.002976	www.gnavi.co.jp/kanto/
4	0.002924	my.gnavi.co.jp/Rating/regist.php?shopid=a275100&shopurl...
5	0.002304	www.nilim.go.jp
6	0.002304	www.icube-t.co.jp
7	0.002304	www.google.co.jp/custom
8	0.001833	www.tsukuba.ad.jp
9	0.001817	www.i-tsukuba.com
10	0.001815	www.ibarakiken.net
11	0.001813	www.e-tsukuba.jp
12	0.000468	rm.gnavi.co.jp/Map/mc_view.php?dr=a275100&c=36...
13	0.000468	rm.gnavi.co.jp/Map/mc_view.php?dr=a275100&c=36...
14	0.000468	rm.gnavi.co.jp/Map/mc_view.php?dr=a275100&c=36...
15	0.000468	rm.gnavi.co.jp/Map/mc_view.php?dr=a275100&c=36...
16	0.000468	rm.gnavi.co.jp/Map/mc_view.php?dr=a275100&c=36...
17	0.000468	rm.gnavi.co.jp/Map/mc_view.php?dr=a275100&t=s
18	0.000468	rm.gnavi.co.jp/Map/mc_view.php?dr=a275100&c=36...
19	0.000468	rm.gnavi.co.jp/Map/mc_view.php?dr=a275100&c=36...
20	0.000468	rm.gnavi.co.jp/Map/mc_view.php?dr=a275100&c=36...

図 8　未ダウンロードの URL に対するランク計算結果（上位 20 件）