中間発表

06T4056N　三沢博章

本日の流れ

- 本研究の題材としている論文の紹介
- 現段階の着手状況
- 最終目標

題材としている論文

題材:「地域ウェブ情報源のためのクローリング手法の提案」

研究所:
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
筑波大学第三群情報学類

著者:
張　健偉　　石川　佳治　　黒川　沙弓　　北川　博之

はじめに

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。

しかし、、、
地域情報のクローリング手法においてはいくつかの問題点があげられる。

地域情報をクローリングする際の問題点

- ページ地域性の判定の難しさ

  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など

  地名が一般名詞として出現する場合:

  例:札幌ラーメンなど
- ウェブ全体でのページ評価の有用性の問題

  ある地域に対して重要な情報を提供していなくても被リンク数が少ないなどの理由から、ウェブでの評判がよくないという場合

提案手法

ユーザがすでに保持している着目地域に関するデータ(実データ)を有効利用し、選択的なクローリングの実現を目指す

実データ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構築する

↓

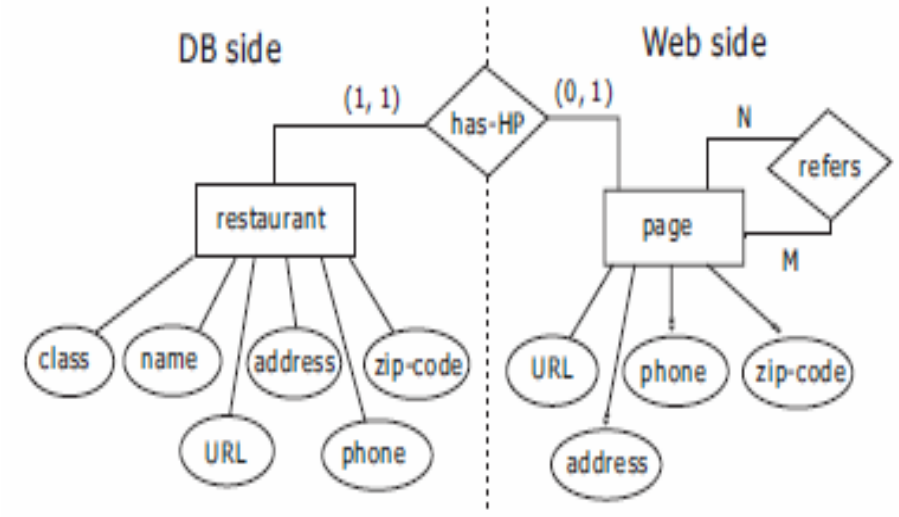
リンク解析を適用し、スコアを計算。

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

データベースの拡張

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



- ユーザーが左図のような、レストラン情報を持っているとする

↓　拡張

- ユーザが提供するデータベースをウェブ上の情報を元に補完、拡充をすることに焦点をおいている。
- 実体関連モデル風に表現したもの。
- 左側が実データを表し、右側がウェブを表す。

リンクの生成

（１）実データのURL属性の値を元にホームページ、および、ホームページから辿ることのできる同一ドメインのページを収集

（２）バックリンクをすべて収集

（３）name属性とphone属性の値をキーワードに検索し関連するページ群を収集

（４）各ノード間のリンクを生成

リンク解析

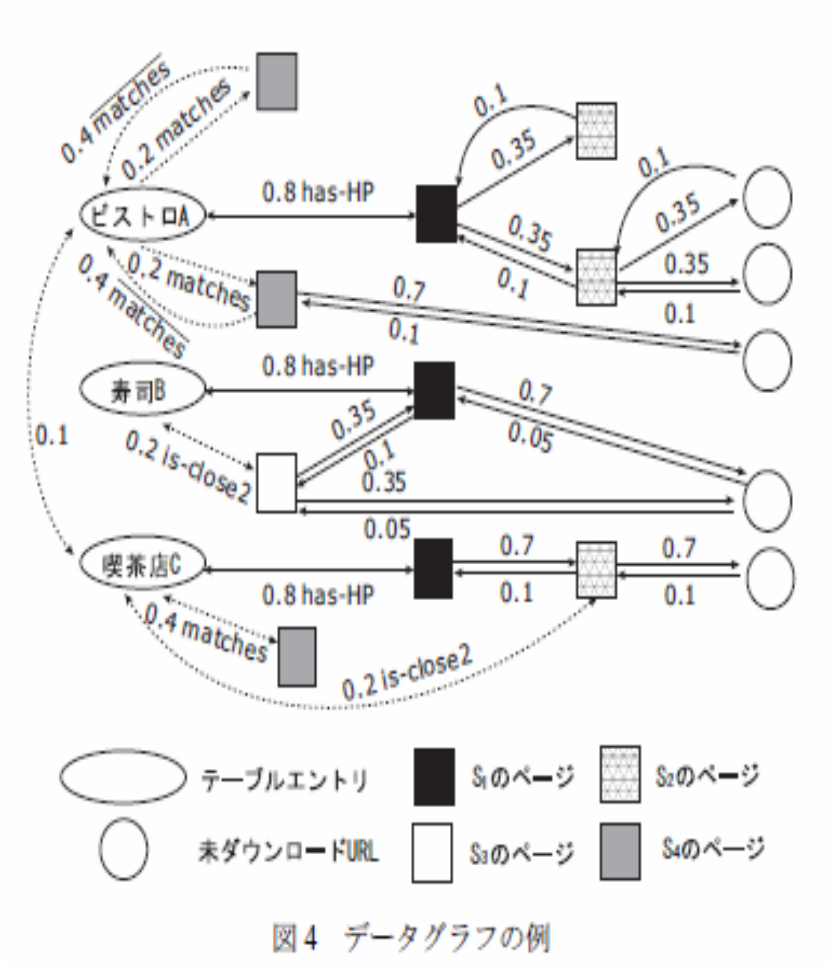


図4　データグラフの例

S_1・・・実データの各行に対しURL属性の値をもとに得たホームページの集合

S_2・・・各ホームページから辿れる同一サイト内のページの集合

S_3・・・各ホームページのバックリンクの集合

S_4・・・店名と電話番号をもとに検索をして得られたページの集合

データグラフ・・・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

スコアの計算

先ほどのデータグラフにおいて各ノードのLocalRank値を計算する。

LocalRank値・・・データベースのエントリと対象地域との関連度を表すスコア

LocalRank値が大きいほど対象地域との関連度が高いことを表す。

この値が最も大きいURLを選んでアクセスすることで地域性の高い情報を選んで、選択的なクローリングを実現できる

この方法によって・・・

例：
ユーザはつくば市の飲食店に興味がある。
ユーザはつくば市の飲食店情報を所持している。
どうせなら、つくば市の地域性のある店に行きたい。
（チェーン店などではなく）

↓

そこで、クローリングを行うと、、、

解析結果

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/

支店が存在しないようなローカルな飲食店はスコアが高く

吉野家といったチェーン店の店は低くなっている

現在の着手状況

・題材としてる論文との相違点

人手で抽出した実データ（つくば市の飲食店の情報）
　を自動で取得

↓

この際、データ源として「ぐるなび」のサイトを利用

現段階では、着目する地域名を入力し、
その地域の飲食店のURLと店名の取得まで終了
(50件まで取得）

実行結果

```
C:¥practice>perl http3.pl つくば
http://r.gnavi.co.jp/e218500/
 - つくばYOUワールド
http://r.gnavi.co.jp/a576600/
 ―地酒・地鶏・鮮魚・宴会―  つくば亭
http://r.gnavi.co.jp/e395300/
 伊太利亜台所 TRENO つくば
http://r.gnavi.co.jp/g220324/
 炭火焼肉トラジ つくば店
・・・
http://r.gnavi.co.jp/a235602/
つくば 茶寮 楽
http://r.gnavi.co.jp/g658000/
鶏料理と串焼き つくば地鶏御膳 わさび家
http://r.gnavi.co.jp/a235600/
つくば 鮨 き羅く
http://r.gnavi.co.jp/b697400/
 串とんぼ つくば店
```

最終目標

- より大きなデータを用いて、どのような結果が得られるか検証する

  実データの取得を人手でなく自動で行うため、データの確保は短時間で可能(?)

- 対象データ（今回は飲食店）を、ほかのカテゴリなどで利用できないか、拡張し実験を行い検証を行う