2012年度　卒 業 論 文

ウェブ文書の構造を利用した
場所名・住所ペアの獲得

2013年3月31日

情報知能システム総合学科
(学籍番号: A9TB2106)

佐藤 貴大

東北大学工学部

概 要

近年見られるようになった、ツイッターを始めとするなマイクロブログの爆発的普及に伴い、位置情報を扱う研究の重要性が上昇してきている。位置情報を推定する上で、文中に含まれるランドマークを特定し、その住所と結びつけることは非常に重要である。しかし、文中に出現し得るランドマークの種類は多様であり、その共通点なども見つけにくいことから、その特定は容易ではない。このため、場所に関する言語資源を大量に獲得することは有用である。

本稿では WEB 文書中に存在する、ランドマークと住所がまとめられた住所一覧ページから、高い精度でランドマークとそれに対応する住所のペアを抽出する手法を提案する。ページ内でのランドマークと住所のペアの間のパスを構築し、得られたパスに制約をかけ、それをもとにしてランドマークと住所のペアの獲得を試みることで高精度の抽出を行う。また、住所一覧ページ間でのパスの共有により、出現パターンの獲得が行えなかったページからの抽出も行う。

実験によりパスに対する制約の効果を確かめ、提案手法での精度の上昇を示す。さらに、ページ間でパスを共有することによる獲得ペア数の上昇とそのときの精度も示す。

目 次

第1章　序論

1.1　背景

近年爆発的な普及を見せたTwitterに代表されるマイクロブログにおいて、人々は日常のあらゆる行動を手軽に日記のようにして投稿することが可能となった。マイクロブログの書き手の多くは一般の個人であり、その内容から、書き手の思想や行動、人々のつながりなどを抽出できる可能性がある。また、リアルタイム性の高い情報発信がなされるという特徴もある。このため、マイクロブログからの情報の抽出、マイニング、書き手の属性推定、個人に対する行動分析などの重要性が高まってきた。

マイクロブログを扱う上で重要な意味をもつ情報の一つに、位置情報がある。例えば書き手の属性推定を行う場合、ユーザの居住地域を推定することは非常に効果的である[1]。居住地域を特定することで、「地元」の人の意見の信頼度を高く考えることや、マーケティングにおいては特定の地域に住むユーザのみへの情報配信などが可能となる[2]。その他にも、地域を限定したトレンド分析や評判分析などにも有用である。また、個人に対する行動分析にも、位置情報は有益である。マイクロブログが投稿された場所（あるいは、マイクロブログ上で示唆している場所）を特定することで、あるユーザがその日訪れた場所、そこでとった行動などを分析する助けとなり得る[3]。

Twitterにおいては各投稿に対してその位置情報としてジオタグが付加されている場合がある。このジオタグは位置情報を利用した研究において非常に有益ではあるが、ジオタグ付きツイートの割合は小さいため、直接この情報を利用できるものもあまり多くはない[4]。このため、ツイートの本文中の単語を利用したユーザの位置推定の研究が多くなされている[4][5][6]。

このようにマイクロブログ上で位置情報を扱う場合において、文中に表現されるお店や学校、交差点のような場所を表す表現の認識と、その場所（住所）の特定は大きな意味を持つ。しかしながら場所を表す表現は多岐にわたり、その表現に規則性のようなものを発見するのは容易ではないため、正規表現での獲得や、文字列の規則性から特定の文字列を判定する系列ラベリング問題としての抽出は難しい。そのような場合、場所に関する言語資源を大量に獲得できていれば、マイクロブログ上の本文と獲得した場所の情報とのマッチングにより、場所の表現を認識して、その場所と住所を結びつける助けになる。

1.2 目的

ウェブ上には人によりランドマークと住所の情報がまとめられた住所一覧ページが多数存在する。これらの多くは人手によりまとめられたもので、ランドマークと住所についての情報がページ内に規則的に配置しまとめられているため、この規則性を獲得できればページ内のランドマークと住所のペアを高精度で獲得できる。また、「東北地方のホテルについてのまとめたページ」、「宮城県内の学習塾についてまとめたページ」、「テレビで取り上げられた有名なお店についてまとめたページ」のように、まとめられる対象やそのくくり方など多岐にわたるため、大量のランドマーク・住所ペアの獲得が見込まれる。

本研究では、この住所一覧ページに着目し、ページの中でのランドマークと住所の記載のパターンを獲得し、どのランドマークとどの住所が対となるものなのかを当て、ランドマーク・住所ペアの抽出を高精度で大量に行うことで、場所に関する言語資源を拡張していくことを目的とする。

住所一覧ページにはランドマークと住所が規則的にまとめられている。しかしその規則性はページ毎に異なるため、各住所一覧ページについてその規則性を見つける必要がある。この規則性をもとにして、各住所一覧ページにおいて、住所に対応しているランドマークはどこに配置されているのかをとらえることでペアの獲得を行う。ページ内でのペアの出現パターンをとらえるため既知のランドマーク・住所ペア（シードデータ）とのマッチングを用いる。住所の検出は正規表現を用いて見つけることが出来る。マッチングにより見つかった住所からランドマークまでのページ内での経路（パス）を用いて、住所を基点としてページをたどることで対応するランドマークを見つけ、ランドマークと住所のペアを獲得する。

ページ内で構築されたパスをもとにしてペアの獲得を行うため、パスの構築の失敗はシステムの精度を大きく下げる要因となる。このため、ペアの獲得の際に用いるパスに制限を加えることで誤ったパスを除外し、これにより精度の向上を図った。

また、パターンをとらえる際にシードデータとのマッチングを用いるため、住所一覧ページにまとめられたのランドマークと住所の中に既知のものが存在していない場合、そのページからの抽出がいっさい行えないという欠点が存在する。この問題を解決するため、本研究では、住所一覧ページの URL のドメインと階層をもとに複数のページをまとめ上げ、その中でパスを共有することで、本来パターンをとらえられていなかったページからの抽出を行う。パスの共有を行っても精度の低下が見られなかったこと、新規獲得ペア数の大きな増加が見られたことを示す。

1.3 構成

本稿は全部で５章からなる。本章に続く２章では住所一覧ページからの場所に関する言語資源拡張の関連研究を紹介する。３章ではベースラインとして用いた手法の説明とその際生じる精度の問題と獲得数の問題を述べ、提案手法について詳しい説明を行う。４章では２つの実験の設定

と評価尺度、結果、およびそれぞれの分析について述べる。５章では本研究で明らかになった点と、今後の課題について述べ、本研究のまとめを行う。

第2章　関連研究

近年、位置情報の重要性は著しく上昇してきている。Twitter などの位置情報を含んだフィードを地球儀にマッピングしてくれるサイトや、マイクロブログを投稿した位置情報を複数人で共有できるアプリなどは位置情報に対する関心の高まりを示す [9][10]。

位置情報を用いたユーザの居住地域推定によって、「地元」の人間の情報の信頼度を高く考え分析を行うことや、得られた居住地域情報を用いて特定の地域に住むユーザのみへの情報配信を可能とする [2]。地域限定の情報配信は企業にとっては情報伝達の効率化とそのコスト低下のメリットが見込め、ユーザにとっては情報の取捨選択の必要性を軽減する。

マイクロブログの一種である Twitter では位置情報としてジオタグが付加される場合が存在する。ジオタグは Twitter においては投稿したときの所在地データを追加し、他ユーザにどこから投稿したのかを知らせるもので、これを利用したユーザの行動パターン推定に関する研究もなされている [3]。しかしジオタグが付加されたツイートの割合は Twitter 全体に対して非常に小さく、Cheng[4] らによれば、全体のわずか 0.42 %ほどのツイートにしかついていない。そのため、ツイート中に出現した単語と、投稿された位置との関係をからユーザの位置推定などの研究がなされている。Cheng ら [4] は、特定の位置との結びつきの強い単語が存在するという考えから、マイクロブログ中の各単語の位置との相関（条件付き確率）をもとにした確率モデルによって、市レベルでのユーザの一推定を行う手法を提案している。Eisenstein ら [6] は、単語と地域の結びつきをもとにして、潜在トピックと地域を一緒に推論するマルチレベル生成モデルを提案している。

場所に関する言語資源の充実は位置情報を効果的に扱う助けとなる。例えば「東北大学工学研究科・工学部」の住所が「宮城県仙台市青葉区荒巻字青葉 6-6-04」であるという情報と「理薬食堂」の住所が「宮城県仙台市青葉区荒巻字青葉 6-3」であるという情報を獲得できれば、文中に「東北大学工学研究科・工学部」という文字列が出現した際、それが 1 つのランドマークを示し、周辺のランドマークに「理薬食堂」が存在していることがわかる。これにより、周辺のお店として「理薬食堂」を推薦することが可能となる。

本研究の先行研究として、村山らによる WEB 上の住所一覧ページをもとに場所に関する言語資源の拡張を行った研究が存在する [7]。この研究では、ランドマーク・住所・電話番号の三つ組の抽出を行っていた。

まず、住所一覧ページから DOM ツリーを作る。DOM ツリーとはウェブページを各要素をノードとする木構造で表現したものである。ウェブページの各要素にはタグが

付けられているため、DOM ツリーのノードはそれぞれタグを持っている。もとのウェブページの構造から、テキストを含むノードも存在する。DOM ツリーにおいて、あるノードから別のノードまでにたどったノードの並びを「パス」と表現する。例えば、図 2.2 において、「022-222-5373」を含むノードから「仙台市青葉区ホテル一覧」までのパスは

$$< \text{td} > \uparrow < \text{tr} > \uparrow < \text{table} > \uparrow < \text{div} > \downarrow$$

のようになる。

住所一覧ページを DOM ツリーとして扱うことで、三つ組のページ内でのまとめられ方をとらえることが出来る。

抽出は、DOM ツリーからのパスの構築と、発見したパスを用いての三つ組獲得にわけられる。

パスの構築では、既知の三つ組のシードデータとのマッチングを行う。まず、最も簡単に見つかると思われる電話番号のマッチングから行う。DOM ツリーの各要素とシードデータの電話番号とを比較し、次に、マッチした電話番号と組になっているランドマークと住所のマッチングを行う。DOM ツリー内に三つ組が発見されたら、電話番号からランドマークと住所へのパスを構築する。

三つ組の獲得では、まず、DOM ツリーから電話番号を正規表現を用いて検出する。検出された電話番号を基点として、構築されたパスをたどって新規のランドマークと住所を発見する。パスをたどることに成功し、新たに三つ組が発見された場合、新規の三つ組として獲得する。

ウェブ上の一覧ページを対象に抽出を行う研究として、ランドマークや住所以外を扱う研究も存在する。

Labsky ら [8] は、自転車の製品情報の抽出のため、オントロジーの知識と、画像の潜在的意味解析を組み合わせる手法を提案している。

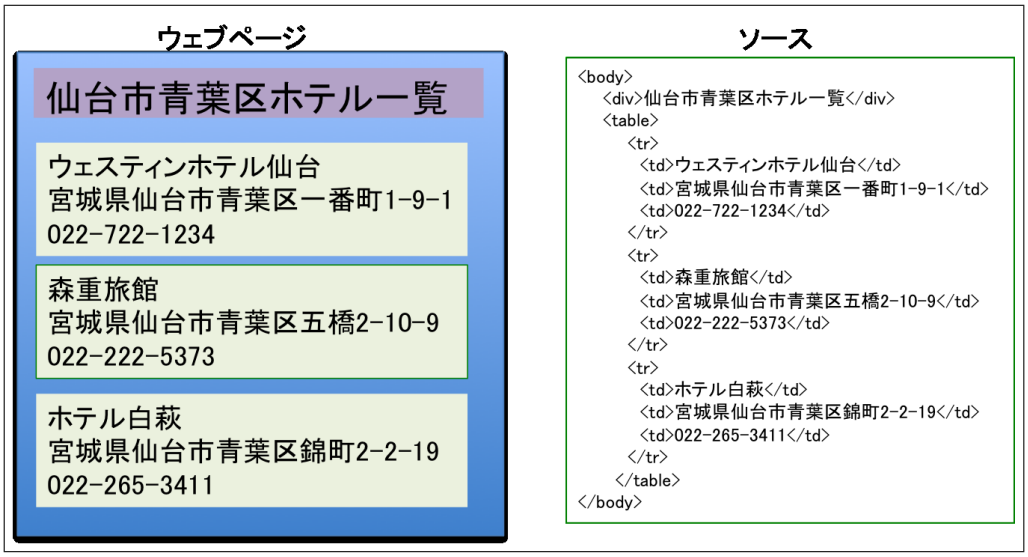

図 2.1: ウェブページ

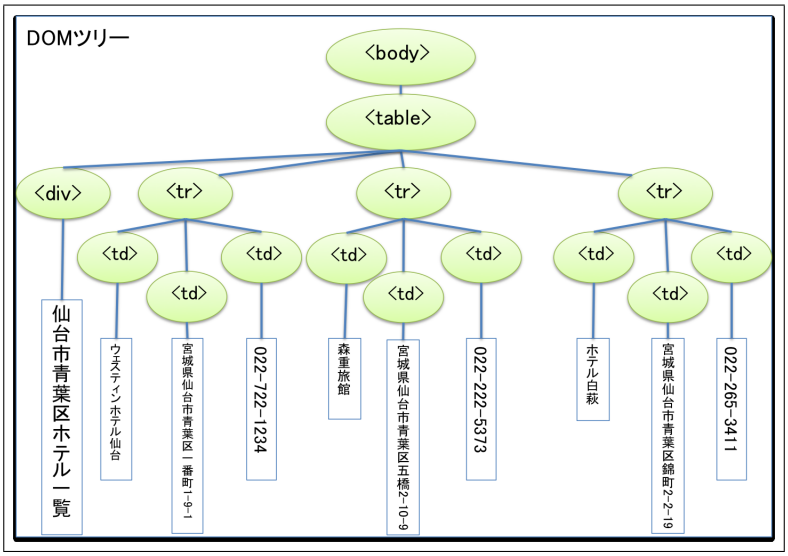

図 2.2: DOM ツリー

第3章　提案手法

２章で述べた、住所一覧ページからのランドマーク・住所・電話番号抽出手法をもとに抽出を行う。先行研究の例では一覧ページ内に電話番号の記載を必要とするが、住所一覧ページにはランドマークと住所がまとめられていても電話番号が記載されていないページも多い。このため、このような電話番号の無いページからの抽出も行うため、ランドマーク・住所ペアの抽出を行う。

本手法は大きく４つの行程を行う。

パスの構築　住所一覧ページから DOM ツリーを作る。作られた DOM ツリーとシードデータとのマッチングから DOM ツリーにおける住所からランドマークまでのパスを構築する

パスへの制限　構築されたパスの中で正しく住所からランドマークまでのパターンをとらえられていないものを除くため、使用するパスに制限を加える。

パスの共有　パスの構築が出来なかったページからのペア獲得を行うため、URL のドメインと階層が同じページ間でパスを共有する。

新規獲得　正規表現により新規住所を見つける。住所からパスをたどり、対応しているランドマークを発見し、新規ペアを獲得する。

以下にそれぞれの行程の詳細を示す。

3.1　パスの構築

住所一覧ページからのパスの構築を行う。まず、ウェブページの構造を扱うため、ページから DOM ツリーを作る。DOM ツリーの各ノードについて、シードデータとのマッチングを行う。全ノードに対してマッチングを行い、ページ内の全ノードについて既知の文字列かどうかを判定する。検出されたすべての既知の文字列に対して、シードデータをもとに正しい組み合せを調べる。既知のランドマークと住所の正しいペアを特定したら、DOM ツリーをたどり、住所からランドマークまでのパスを構築する。

3.2　パスへの制限

住所一覧ページから構築されたパスの中には図 3.2 に示すように、正しくないものが存在する。これは、一つのページ内に同一のランドマークや住所が複数回出現した場合に起きる。例では、「ビジネスホテル花木」と「ホテル ハナキ (HANAKI)」は同一のホテルであり、住所も同じであるが、例のページでは表記の違いから別のランドマークとして２回出現している。このとき、「ビジネスホテル花木」と「宮城県多賀城市八幡 4-8-33」が既知のペアであるならば、「ホテル ハナキ

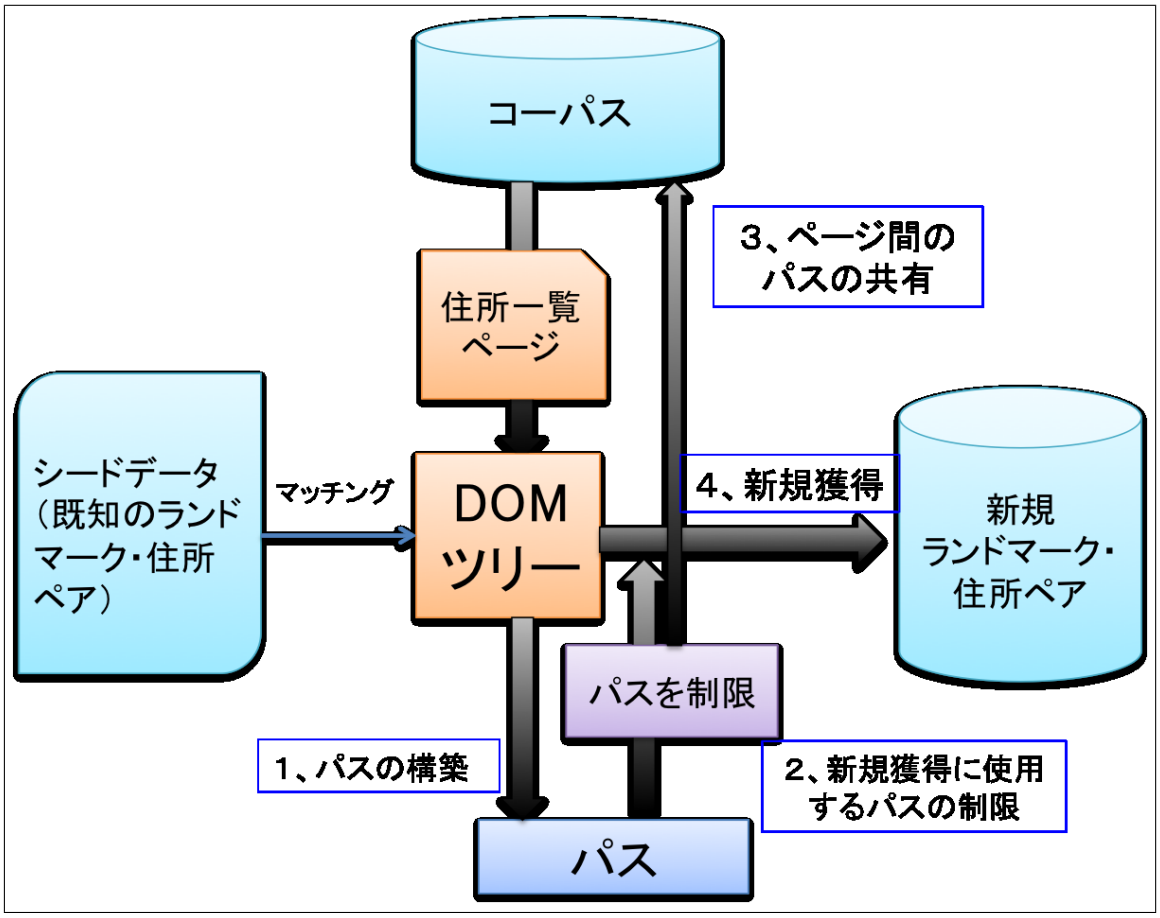
コーパス
住所一覧
ページ
3、ページ間の
パスの共有
シードデータ
（既知のランド
マーク・住所
ペア）
マッチング
DOM
ツリー
4、新規獲得
新規
ランドマーク・
住所ペア
パスを制限
1、パスの構築
2、新規獲得に使用
するパスの制限
パス

図 3.1: 手法概要

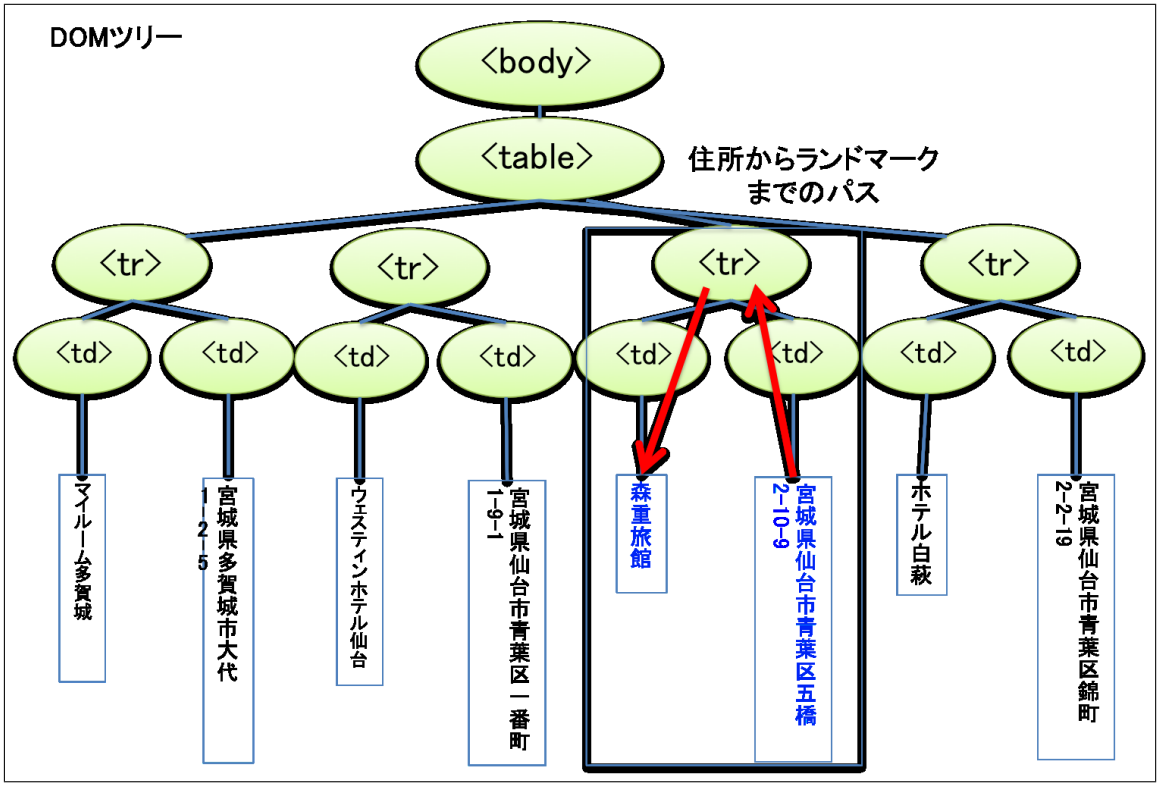

図 3.2: 住所一覧ページからのパスの構築

(HANAKI)」のものとして出現した住所からも「ビジネスホテル花木」との間にパスが構築されてしまう。このとき構築された２種類のパスをもととして抽出を行った結果が図 3.3 である。新規ペアは構築したパスに対して複数獲得されるので、誤ったパスは精度を大きく下げる要因となってしまう。

その他のページとは異なり、住所一覧ページでは、複数のランドマークと住所をまとめて記述する都合上、同じ実体について何度も言及することは少ない。しかしながら、それでもランドマークや住所が複数回出現する例が存在する。例えば、次のような場合があげられる。

同じ建物への表記の揺れから２度取り上げられた場合

図 3.2 に示した例がこれに該当する。住所一覧ページを作る際に、同じ建物について異なる呼ばれ方がなされていたため、ページ内で重複が起きてしまったケースや、ページ編集者が意図的に２度あげたものなどが考えられる。

同一の住所に複数のランドマークが存在する場合

住所に存在する建物がビルである場合などがこれに該当する。例えば「ヨシミ キッチン」と「マンゴツリーカフェ 仙台パルコ店」は共に仙台パルコのビル内に存在しているため、共にその住所は「宮城県仙台市青葉区中央１丁目２－３」である。

時間経過とともにランドマークが移り変わった場合　店の移転や閉店、買収などによって、同じ住所でもランドマークが変化する場合がある。特に東北地方では、2011 年の震災により多くの店舗が移転、閉店したため、同一の住所でも時期によってランドマークが異なっている例

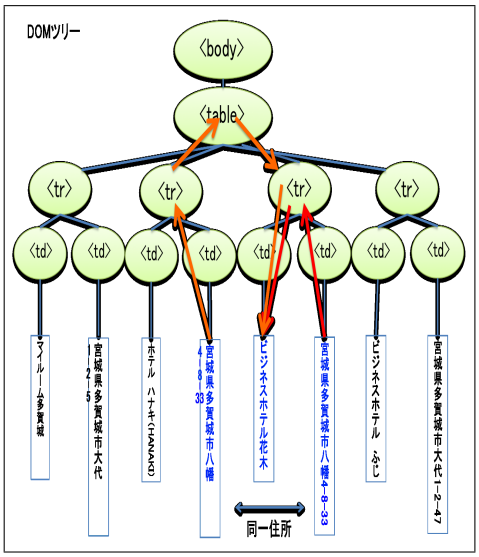

図 3.3: 誤ったパスの構築

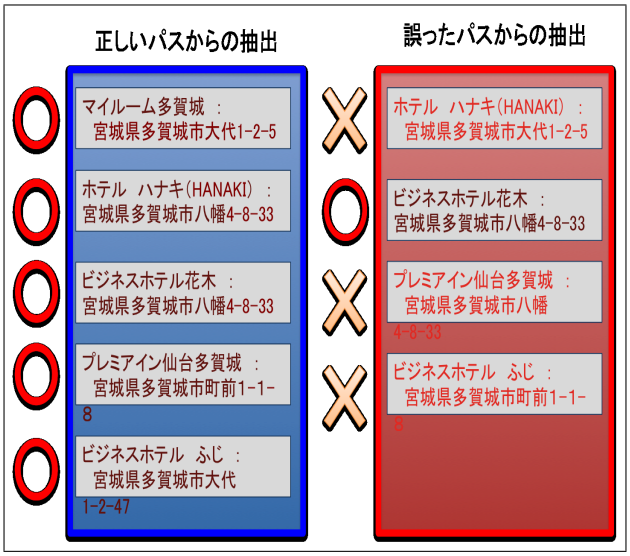

図 3.4: 抽出結果

が多い。

以上のことから、高精度のランドマーク・住所ペアの抽出を行う上で、正しいパスの選択は必要である。本研究では、ランドマーク・住所ペアの獲得に用いるパスに対する制限として以下の２種類について比較をおこなった。

- 頻度による制限
  - パスの出現頻度に対して制限をかける。
  - 構築されたそれぞれのパスについて、ページ内で最も高頻度で出現したパスを用いる。
- 距離による制限
  - ランドマークと住所の間のパスの長さ制限をかける。
  - ランドマークから住所に至るまでにたどったノードの数を距離として、ページ内で最も距離の短いパスを用いる。

２種類の制限の優劣は自明でないため、実験により比較を行う。より高い精度での抽出を行った制限を採用する。

3.3 ページ間のパス共有

本手法では住所一覧ページからパスを構築する際に、既知のランドマーク・住所ペアとのマッチングを必要とするため、１つも既知のペアが存在していないページからは抽出が行えないという欠点がある。また、たとえ既知のペアに相当するランドマークが存在しても、表記の揺れによりパスが構築できない例も存在する。例えば、図 3.2 の例において、「ホテル ハナキ (HANAKI)」と「ビジネスホテル花木」は同じホテルの名前である。例では「ビジネスホテル花木」が既知の

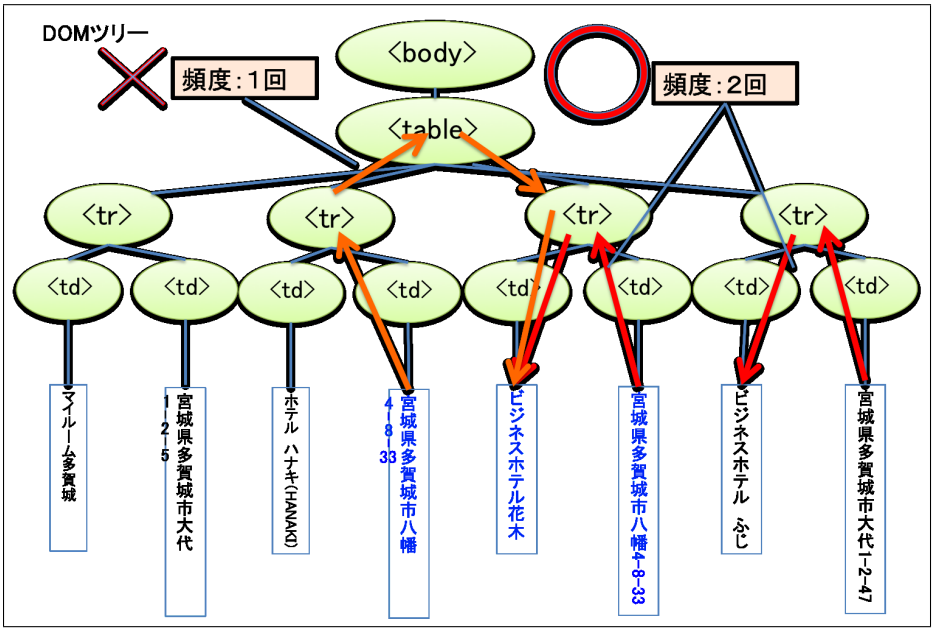

図 3.5: 頻度による制限

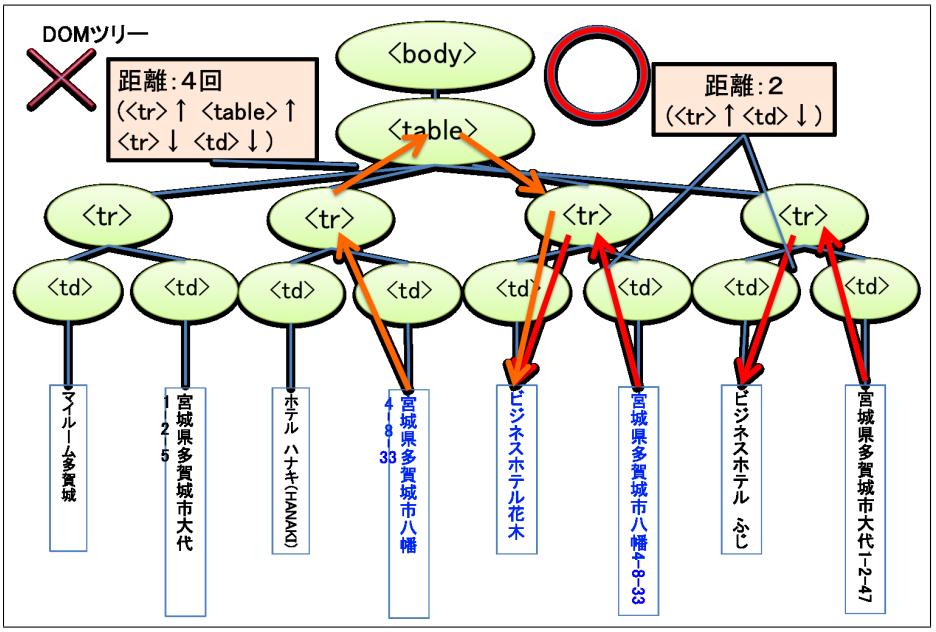

図 3.6: 距離による制限

図 3.7: URL を用いたドメイン、階層が同じページのグループ化の例

ものであったためパスを構築することが出来たが、同じランドマークを示している、「ホテル ハナキ (HANAKI)」は、既知のものと表記が異なるためパスの構築には使えない。

このような理由からパスの構築が出来ず、ランドマーク・住所ペアの抽出ができなかったページにも対応させ、新規獲得ペア数を増加させるために、本研究では URL のドメインと階層が同じページ間でのパスの共有を行う。

パスの構築の行程においてパスを構築できたページについて、URL のドメインと階層を確認する。同ドメイン同階層の住所一覧ページであれば住所からランドマークまでのパスが同じであるという仮定に基づき、コーパス内でドメインと階層が同じ住所一覧ページからのペア獲得を行う際、たとえそのページ内でパスを構築することが出来なくても、ドメインと階層が同じ他のページで構築されたパスを利用する。

本来パスを構築できていなかったページに新たにパスを与えていくことになるため、これによる精度の低下が見られないか実験を行う。また、目的である新規獲得ペア数の増加がどれほど達成されているかを確認する。

3.4 パスに基づく新規ペア獲得

各住所一覧ページについて、ページ毎に構築または共有したパスをもとに新規のランドマーク・住所ペアの獲得を行う。

住所一覧ページから作られた DOM ツリーの各ノードについて、正規表現を用いて住所を検出する。検出された住所毎にパスをたどって対となるランドマークを見つける。ツリーに対応するノードが存在せずパスをたどれない場合や、パスをたどった先のノードが文字列を持たない場合はその住所からの獲得を行わないものとする。正しく対となるランドマークを見つけることが出来たとき、その住所とランドマークを新規ペアとして獲得する。

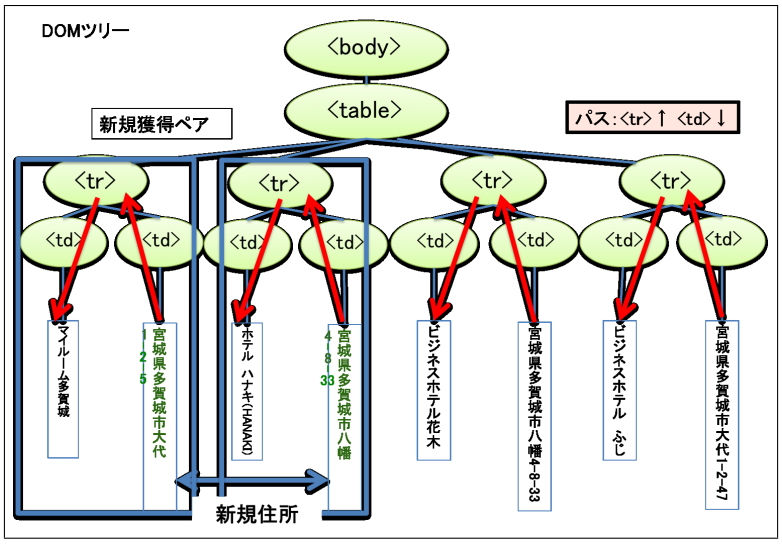
DOMツリー
<body>
<table>
新規獲得ペア
パス：<tr>↑ <td>↓
<tr>
<td>
マイルーム多賀城
宮城県多賀城市大代
1−2−5
ホテル ハナキ(HANAKI)
宮城県多賀城市八幡
4−8−33
ビジネスホテル花木
宮城県多賀城市八幡4−8−33
ビジネスホテル ふじ
宮城県多賀城市大代1−2−47
新規住所

図 3.8: 新規ペアの獲得

第4章　実験

本章では、正しいパスを選ぶための制約としてパスの頻度と距離のどちらについての制約をかけるべきかの選択と、ページ間でパスを共有した際の精度と新規獲得ペア数の確認の２つの実験について、それぞれの実験設定、評価尺度、実験結果を記し、それらに対する分析を行う。比較対象として、パスに対しての制限並びにページ間のパスの共有を行っていないものについての評価も行う。これにより、パスを制限することによる精度の向上と、抽出の精度が実用の耐えうるものであるかどうかを調べる。

4.1　パスに対する効果的な制限の比較

3.2 節で述べたように、高い精度での抽出を行うため、構築されたパスから正しいもののみを選択する必要がある。本実験ではパスに対する制限として適切なものを確かめるため、それぞれの制限に対する実験結果により比較を行う。

4.1.1　実験設定

実験にはウェブから収集された約 6700 万ページの日本語ページを持つ ClueWeb09 の資源のうち約 700 万ページから、２種類以上のルートからのパスが同じ住所を含む住所一覧ページ約２万ページを用いた。

シードデータとしては Yahoo!ロコから抽出された宮城県内のランドマーク・住所ペア約 10 万対[1]を与えた。

比較対象として、以下の３通りの実験を行った。

- パスに対する制限を加えない (構築したパスをすべて適用する)
- 最も出現頻度が高いパスを選択する
- 最もランドマークと住所の距離が短くなるパスを選択する

4.1.2　評価尺度

評価は適合率, 網羅率 (全体), 網羅率 (新規), 全新規獲得ペア数の 4 種類により行う。適合率、網羅率 (全体), 網羅率 (新規) の定義はそれぞれ以下の式 (4.1), 式 (4.2), 式 (4.3) に示す。

$$\text{適合率} = \frac{\text{宮城県内正解ペア数}}{\text{抽出された宮城県内のランドマーク・住所ペア数}} \tag{4.1}$$

[1] コーパス内の WEB ページの収集年 2009 年に対しシードデータの収集年は 2012 年である

$$\text{網羅率 (全体)} = \frac{\text{宮城県内正解ペア数}}{\text{コーパス内の全宮城県内住所件数}} \tag{4.2}$$

$$\text{網羅率 (新規)} = \frac{\text{宮城県内新規正解ペア数}}{\text{コーパス内の未知の宮城県内住所件数}} \tag{4.3}$$

獲得されたランドマーク・住所ペアの正否判定については、コーパスの収集年の都合上、自動判定が難しかったため、宮城県内のものに限定した上で人手で判断した。

4.1.3 実験結果

実験の結果を表 4.1 に示す。

表 4.1: パスに対する制限の比較

パスへの制限	適合率	網羅率 (全体)	網羅率 (新規)	新規獲得ペア数 (対)
なし	0.743	0.252	0.196	2,747
頻度最大	0.916	0.251	0.195	2,435
距離最短	0.997	0.247	0.191	1,963

パスの制限を行わなかったときと比較し制限を加えることで適合率が約 0.20 ほど上昇した。新規獲得ペア数に関しては、出現頻度が最大となるパスのみを選ぶ制限を加えた事により約 300 対ほど減少した。ランドマークと住所の距離を最短とするパスを選ぶ制限を加えたことにより約 800 対ほど減少した。しかし全宮城県内住所件数に対する宮城県内の正解ペア数である網羅率の低下が見られないことと、適合率の大幅な上昇から、減少した新規獲得ペアの大部分は誤ったパスにより抽出された正しくないランドマーク・住所ペアであると考えられる。

4.1.4 分析

実験の結果、新規獲得ペア数の観点からは、パスに制限を加えたときがのほうが獲得できるペア数は少なかった。しかし制限を加えなかったものと比較したときに、獲得した新規ペア数は 3 割近く減少していながら網羅率の低下がわずかであったこと、適合率は 0.25 も上昇し、ほぼすべての新規獲得ペアが正解であったことから、この獲得数の低下は誤った抽出を防ぐことの出来た結果であり、問題はないといえる。

また、本実験設定でのパスに対する制限は、ランドマークと住所の距離を最短にするパスの選択のほうが優れていると判断できる。これは、ランドマークと住所の距離を最短にするパスの選択のほうが本実験設定においては優れていたこととなる。この原因として考えられるのが、ページ内における正解パスの出現回数の低さである。実験においてパスの構築に利用された既知のランドマーク・住所ペアの数は 150 対のみで、パスを構築できたページ数はコーパスの約 2 万ページ中わずか 63 ページであった。このことから、パスを構築できたページにおいても、その構築に利用された既知のペアは 1 ページ当たり約 3 対程度であったことがわかる。これに対して頻度を用いて制限を加えた結果、誤ったパスと同頻度でしか正しいパスが構築されなかったページが出現し、これにより距離を用いた制限よりも適合率が低かったと考えられる。

パスの構築に利用された既知のランドマーク・住所ペアの数が少ない理由としては、シードデータの作成とコーパスの収集年に差があることも 1 つの要因であると考えられる。コーパス中のウェブページが収集された 2009 年からシードデータが作成された 2012 年までの間に発生した東日本大震災の影響により移転・閉店が生じたため、コーパス収集時には存在していたがシードデータ作成時には存在していなかったランドマークが存在したため、マッチングが取れずパスが構築できなかったページが存在していたことの影響が生じていたと考えられる。

また、別の要因としては、ランドマークの表記揺れが考えれる。1 つのランドマークに対する呼び名は必ずしも 1 通りに定まらないため、同じ対象を示していてもマッチングをとることが出来なかったものが存在していた。

距離を用いたパスへの制限が 0.997 もの適合率を達成した原因としては、誤ったパスが構築されるときは図 3.2 で示したようにして一覧ページ内の 1 つのランドマークに関するまとまり (レコード) をまたいでしまう。多くのページにおいてレコードをまたぐ場合のパスはレコード内部のパスの距離よりも長くなるため、距離を用いた制限はこれを防ぐ効果を持ったと考えられる。

パスに対して距離を用いた制限を与えることで適合率は十分実用可能なものとなったといえる。宮城県内では誤って抽出されてきた件数はわずか 1 件のみでこれは下記に示す通り、1 つのランドマーク名が複数行にわたって記載されていたためであった。具体的には「ビジネスホテル ニューシャトー原町」が HTML のタグにおいて「ビジネスホテル」と「ニューシャトー原町」の二つに分かれていたため前半の「ビジネスホテル」のみをランドマーク名として抽出してしまった。このような例においては人間が確認することで誤った抽出であると判断することが可能であると考えられる。

適合率に関しては十分であるといえるが、住所一覧ページ 2 万ページに対して新規獲得ペア数が約 2,000 対であった。これは先述した通り、パスを利用できたページ数が 63 ページしかなかったことが原因であると考えられる。この問題を解決するため、次節にてパスを構築できたページと共通のパスを利用できるページをまとめパスを共有する実験を行う。

4.2 ページ間のパス共有

4.1.4 節でも述べた通り、パスを構築出来るページ数が少ないため新規獲得ペア数が 2,000 対程度となってしまった。ランドマーク・住所ペアの大量獲得のためには、扱えるページ数を増やす必要がある。

シードデータとのマッチングによるパスの構築が行えなかったことによりパスを構築できず、抽出を行えなかったページが多く見られた。このようなページからの抽出を行うことで、扱うことの出来るページ数は増大するはずである。

この問題を解決するため、住所一覧ページ間でのパスの共有を行った。実験によりその際の新規獲得ペア数の増加と適合率の変動を調べる。

4.2.1 実験設定

実験に用いるコーパス及びシードデータは 4.1 節で用いたものと同様のものを用いる。パスに対する制限は、4.1 節で言及した、最もランドマークと住所の距離が短くなるパスの選択を用いる。

比較対象として、以下の 3 通りの実験を行った。

- パスに対する制限及びページ間のパスの共有を行わない
- パスに対し制限を加えるがページ間のパスの共有は行わない
- パスに対し制限を加え、ページ間でパスを共有する

4.2.2 評価尺度

評価は 4.1 節の実験と同様に適合率 (Precision), 網羅率 (Coverage), 新規獲得ペア数の 3 種類により行う。

4.2.3 実験結果

実験の結果を表 4.1 に示す。

表 4.2: パスの共有

パスへの制約・パスの共有	適合率	網羅率 (全体)	網羅率 (新規)	新規獲得ペア数 (対)
制約：なし、　　共有：なし	0.743	0.252	0.196	2,747
制約：距離最短、共有：なし	0.997	0.247	0.191	1,963
制約：距離最短、共有：あり	0.997	0.320	0.269	20,936

パスに対してランドマークと住所の距離を最短とするパスを選択する制約をかけたとき、ページ間でパスを共有しなかった結果と、パスを共有した結果を比較すると、適合率を下げることなく網羅率を 0.07 上昇させ、また、新規獲得ペア数に関しては約 10 倍の増加が見られた。

また、パスに対して制約を加えず、ページ間のパスを共有も行っていない結果と比較すると、適合率で 0.25、網羅率で 0.07、新規獲得数で約 17,000 対もの上昇を確認できた。

このことから、パスの共有を用いることで対象とするページ数を大きく増加させ、より多くのランドマーク・住所ペアの獲得が可能となった。

4.2.4 分析

実験により、本実験設定において、URL を利用してドメインと階層が同じページ同士でパスを共有する手法は有用であるといえる。今回ページ間でのパスの共有に利用した情報は、ドメインと階層のみであるが、対象としているページが住所一覧ページであることから、ランドマークについてまとめられたドメインと階層の同じページ同士の構造が似通っているといえる。

今回の実験でランドマーク・住所ペア抽出に利用可能なパスを持っていたページ数であるが、パスを共有しなかったときの 63 ページに対し、5,115 ページにまで上昇していた。つまり、利用可能なページ数が約 80 倍に増加し、新規獲得ペア数が 10 倍に増加したこととなる。ページ数の増加に対して新規獲得ペア数の上昇が小さいようにも思えるが、これは、パスの共有を行う際に URL

のドメインと階層の一致を利用しているため、抽出できるランドマーク・住所ペアの種類がパスを共有したページ間でかぶっているためと考えられる。例えば、パスを構築できたページがホテルについてまとめられたページであるとき、ドメインと階層が同じ住所一覧ページは、同様にホテルについてまとめられたものが多いと考えられる。これらのページ間でまとめられているランドマークの重複がページ数の増加に対して獲得したペア数の増加が少なかった原因であると考えられる。

今回ページ内でパスを構築することが出来た 63 ページの内、ドメインの違うものが全部で 35 種類存在していた。全コーパス内には 3,065 種類の異なるドメインを持つページが存在し、また、このうちコーパス内に 10 回以上出現したものは 271 種類あった。

今回利用できたページ数は約 5,000 ページであり、これはパスの共有を行う以前に扱えたページ数がわずか 63 ページであったことから、大幅な改善であるといえるが、しかし、コーパス全体が約 2 万ページであったことを考えると、約 1/4 ほどしか活用できていないこととなる。コーパスに含まれる残り約 3/4 の住所一覧ページからのランドマーク・住所ペアの抽出により更なる大量抽出が見込まれる。

第5章　まとめ

本稿では獲得したパスに対する制約とページ間のパス共有による、WEB 中の住所一覧ページからのランドマーク・住所ペアの高精度な大量抽出の手法を提案した。実験の結果本手法において、パスに対して加えるべき制約はパスの長さに基づく最短パス選択であり、さらにパスの共有により、高精度かつ大量のランドマーク・住所ペアの獲得が可能であった。実験により得られた精度は実用にたえられるものであり、また、新規獲得数も大きく向上したといえる。

しかしながら、4.2.4 節でも述べた通り、パスを構築できたページ数はコーパス全体のうちのごく少数で、URL のドメインと階層に基づくパスの共有を行ってもまだなお利用可能なパスを持たなかったことにより抽出が行えなかったページが多く存在する。本稿で提案した手法では、パスの共有を行う際に URL の情報の中のドメインと階層の２つの要素のみにしか着目していない。このため、実際にはパスの共有が可能であるはずの多くのページを見落とす危険性が存在する。

今後の課題としては、URL のみではなく、内部の構造からの類似度の比較などによるパターンの共有や、シードデータの検討によるパス構築可能ページの増加、及び、シードデータとのマッチングを必要としないシステムを考案していくことが求められる。

謝 辞

本研究を進めるにあたり、ご指導を頂いた乾健太郎教授、岡崎直観准教授に感謝いたします。
日常の議論を通じて多くの知識や示唆を頂いた乾・岡崎研究室の皆様に感謝いたします。

参 考 文 献

[1] 松尾 豊, 岡崎 直観, 中村 嘉志, 西村 拓一, 橋田 浩一, 中島 秀之. 位置履歴からのユーザ属性の推定. 情報処理学会論文誌 48(6), 2106-2117, 2007-06-15.

[2] 奥村 学. マイクロブログマイニングの現在. 2012.

[3] 酒巻 智宏, 岩井 将行, 瀬崎 薫. マイクロブログのジオタグを用いたユーザの行動パターンの推定に関する研究. 電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション 110(400), 37-42, 2011-01-20.

[4] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: A content-based approach to geo-locating twitter users. In CIKM ’10, 2010.

[5] 伊川 洋平, 榎 美紀, 立堀 道昭. マイクロブログのメッセージを用いた発信場所推定. DEIM Forum 2012 F7-2. 2012.

[6] Jacob Eisenstein, Brendan O’Connor, Noah A. Smith, and Eric P. Xing. A latent variable model for geographic lexical variation. In EMNLP 2010, 2010.

[7] 村山 紀文, 南野 朋之, 奥村 学. 見たデータ付与のための住所録自動生成. 情報処理学会研究報告. 自然言語処理研究会報告 2004(73), 41-47, 2004-07-15.

[8] Martin Labsky, Vojtech Svatek, Pavel Praks, Ondrej Svab.Information extraction from HTML product catalogues: coupling quantitative and knowledge-based approaches. Dagstuhl Seminar on Machine Learning for the Semantic Web, 2005.

[9] Twittearth.GeocodEarth.URL: http://www.geocodearth.com/, 2012

[10] BASCULE GO!. Pelo. URL : http://www.pelo.jp/, 2012

[11] 西崎 剛司, 奥地 健太, 服部 文夫. 域限定性を考慮した情報推薦における語句抽出の傾向分析とノイズの除去. DEIM Forum 2011 B1-6, 2011.

[12] YAHOO!JAPAN. Yahoo!ロコ.URL : loco.yahoo.co.jp/, 2012