ジオパースによる Web からの空間コンテンツ獲得

相良 毅† 有川 正俊‡

†東京大学生産技術研究所 〒153-8505 東京都目黒区駒場 4-6-1
‡東京大学空間情報科学研究センター 〒153-8904 東京都目黒区駒場 4-6-1
E-mail: †sagara@iis.u-tokyo.ac.jp, ‡arikawa@csis.u-tokyo.ac.jp

あらまし GPS 付き携帯の登場など，LBS(Location Based Service)を実現するための技術は整った．人ナビはその主要なサービスであり，すでに実用化されているものもあるが，整備コストが高いことによるコンテンツの絶対量不足が問題となっている．Web は新鮮な空間コンテンツの有力な情報源であり，Web ページをテキスト解析することにより空間コンテンツとして利用するという研究が行われているが，十分な質と量のコンテンツを短時間で収集する手法の開発とページの内容理解が課題である．本論文では，タウンページを辞書として用いることで Web から特定地域の特定業種に関する空間コンテンツを効率よく獲得し，HTML のタグ構造を利用してコンテンツの検証を行うことで品質を高める手法を示す．

キーワード Web とインターネット，e-service，情報検索，地理情報システム

Spatial Content Retrieval with Geoparsing Web Documents

Takeshi SAGARA† Masatoshi ARIKAWA‡

†Institute of Industrial Science, The University of Tokyo 4-6-1 Komaba, Meguro-ku, Tokyo, 153-8505 Japan
‡Center for Spatial Information Science at the University of Tokyo 4-6-1 Komaba, Meguro-ku, Tokyo, 153-8904 Japan
E-mail: †sagara@iis.u-tokyo.ac.jp, ‡arikawa@csis.u-tokyo.ac.jp

Abstract The shortage of fresh spatial contents becomes one of the biggest problem for location based services, such as human/car navigation and restaurant recommendation. The Web is thought as a promising source of spatial contents so that some research projects had been reported which try to collect and geoparse web pages to exploit, however, it is still difficult to retrieve web pages with high quality and quantity in short period. In this paper, we discuss a way to retrieve spatial contents from the Web efficiently, and an algorithm to verify them, using yellow page as a dictionary.

Keywords Web, Geospatial Information Retrieval, GIS, geoparse, geocoding

1. はじめに

2001 年末には GPS(Global Positioning System)搭載の携帯電話が登場し，常に位置を意識しながら生活することが現実的になってきた．しかし現在のところ，自位置付近の地図の閲覧や簡単な道案内ができるといったアプリケーションが用意されている程度で，位置情報に基づいた情報提供サービス (Location Based Service, 以下 LBS) が十分に提供されているとは言い難い．ハードウェアや空間検索手法，地図データなど，LBS を提供する上で必要な技術や環境は既に整っていることから，われわれは，LBS が普及しないのは提供される情報が質・量ともに不十分であることが原因の一つであると考えている．そこで，このような「位置をキーとして管理される情報」を「空間コンテンツ」と呼び，各種情報源から空間コンテンツを獲得する手法の開発を行っている．

空間コンテンツの有力な情報源として Web が挙げられる．Web 上には企業が設置している Web ページ(以下，単にページという)，情報提供企業または有志個人によるレストランや不動産物件等の Web 情報サイト，掲示板サイトへの書き込みなど，空間コンテンツが豊富に存在している．これらの情報を効果的に活用することができれば，日常生活をより豊かで便利にすることができる[1]．そのためには，Web 上の空間コンテンツを効率的に収集し，そのうち品質の高いページを選別し，業種や名称，地理的な条件によって検索できる情報検索システムが必要である．このような情報検索システムを，本稿では「空間サーチエンジン」と呼ぶ．

空間サーチエンジンは，これまで地名や住所などの場所を示す表記（以下，場所表記という）を含むページを収集し，テキスト処理によって場所表記を抽出(この処理はジオパース,geoparse と呼ばれる)，地球上の位置に変換した上で空間インデックスを持たせ，位置とキーワードによって検索することで実現されてきた[2]．しかし，この手法は全国のさまざまな内容のページを均等に収集するため，ある地域やテーマについて

十分な量のページを収集するには非常に長い時間がかかることなどが原因で，実用的なレベルには達していない．そこで本研究では，タウンページ情報を用いて特定地域・特定業種に関する有用なページを効率的に収集する手法と，収集したページの HTML 構造を解析して順位付けを行う手法を開発し，より実用性の高い空間サーチエンジンを構築した．

以下，2 節では空間サーチエンジンの構築に関連する先行研究をまとめ，3 節で問題を明確にする．4 節で Web から空間コンテンツを収集する手法を，5 節で収集したページの検証と順位付けを行う手法を説明する．6 節で提案手法の評価と実装例を示し，最後に 7 節でまとめと今後の課題を示す．

2. 関連研究

本研究に関連する研究としては，(1) Web から場所表記を含むページを選択的に収集する手法，(2) 収集したページから場所表記を抽出する手法，(3)場所表記に対応する文章中の領域を決定する手法，(4)場所表記を座標値に変換する手法などが挙げられる．

(1) 場所表記を含む Web ページの選択的収集手法

あるページが空間コンテンツであるか否かを判断するにはページの内容を調べる必要があり，比較的負荷の高い処理を行わなければならない．しかし Web 上には膨大な数のサイトやページが存在しているため，総当たりでページを収集しても，十分な割合の空間コンテンツを収集するには長い時間がかかってしまう．そこで，何らかの知見に基づいて Web から空間コンテンツを優先的に，かつ効率よく収集する手法が必要である．この課題に取り組んだ研究としては，リンク文字列に着目する手法[3]と，ドメイン名から公的な地域サイトを発見する手法[4]がある．

(2) 自然言語文章から住所表記を抽出する手法

場所表記のほか，組織名・人名・時間表現・数値表現など，利用者にとって検索のキーとなる情報を固有表現と呼ぶ．固有表現をページのような自然言語文章から抽出する手法は，情報検索の分野で長年に渡り研究が行われている[5]．固有表現のうち人名や地名で特に問題となるのは，日本語特有の問題である複合語と多義語の存在である．たとえば，「札幌ラーメン」は 1 語の複合語（組織名）だが，「札幌（地名）／ラーメン（名詞・一般）」に分解されてしまうことがあり，誤抽出の原因となる．文献[6]では，形態素解析と固有表現辞書，およびパターンマッチルールを用いて，複合語や多義語を高い精度で抽出する手法を示した．

(3) 場所表記に対応する文章中の領域決定手法

あるキーワード，たとえば「ラーメン屋」について検索を行い，見つかったページ内に場所表記が含まれていたとしても，その表記がラーメン屋のものであるとは限らない．実際に検索エンジンを利用して検索した結果を見ると，レビュアーの住所や，ラーメン屋に行く前に立ち寄った居酒屋の住所であるといった場合がある．そのため，文章中に場所表記が現れた場合，その表記が指し示している文章中の領域を決定する必要がある．

このような領域抽出は知的な内容の理解を伴うため，一般にはきわめて困難である．しかし文献[7]では HTML のタグ構造を利用して，<title></title>に囲まれる部分に登場したキーワードはページ全体を支配し，<h1></h1>などの見出しに囲まれる部分に登場したキーワードは次に同レベル以上の見出しが現れるまでを支配する，といったルールを定め，領域抽出を行った．この手法により，同一領域内に存在する店舗等の名所と電話番号，住所の対応付けを行い，Web から自動的に住所録を作成した．

(4) 場所表記を座標値に変換する手法

場所表記を対応する座標値（経緯度など）に変換する手法は，アドレスマッチングまたはジオコーディングと呼ばれている．文献[3]では町字レベルまでの住所表記に対応する地名データベースを作成し，対応する空間領域を求める手法を利用している．われわれのグループでも任意の階層構造を持つ地名を参照する手法を開発し，山・湖沼・河川などの自然地名や，号レベルの住所や建物名など詳細な場所表記も変換可能なシステムを開発し，Web 上でアドレスマッチングサービスとして公開している[8]．

3. 空間コンテンツ収集の問題

空間サーチエンジンの主な用途は，ある場所周辺の店舗や施設等の情報を検索することである．店舗・施設情報の検索には，あらかじめ目的の店舗が決まっている場合と，大まかな利用目的しか決まっていない場合の２通りがある．たとえば待ち合わせ場所の A という喫茶店を探す場合は前者に該当し，知らない街で昼食をとるためのレストランを探す場合は後者に該当する．前者の場合，場所による検索を行わなくても喫茶店の名称などで精度の高い検索を行うことができるので，ここでは後者の場合のみを対象として，問題を次のように定義する．

Q. 空間コンテンツ収集

位置 p 周辺の，業種 c に関連するページ集合 W を列挙せよ．

場所による検索手段のない Web 空間に対してこのような空間検索を行うために，これまではサーチロボ

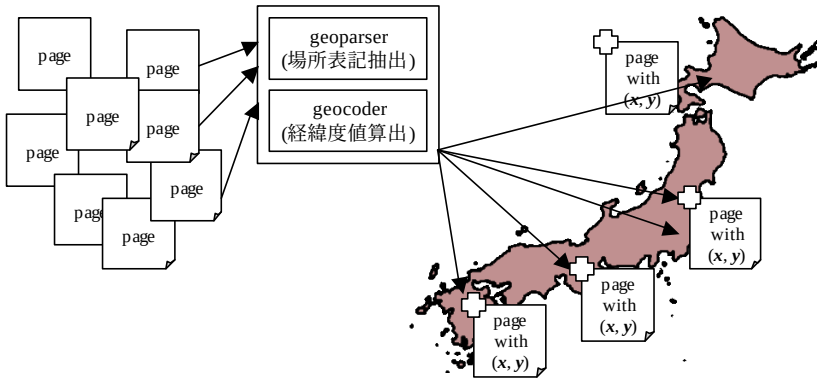

図 1 既存の空間コンテンツ収集手法

ットを用いてあらかじめ大量のページを収集し，収集したページに含まれる場所表記をジオパースして空間的な位置として与えておくという方法をとっている(図 1). これは一般的なキーワードによる検索サービスを提供するサーチエンジンが，まずページを収集し，次にキーワード辞書を作成するのと同じアプローチである．しかし，空間コンテンツは地域限定の情報であるという性質上，リンク関係が少なく，サイト内のリンク階層を深いレベルまで収集しないと重要な空間コンテンツを含むページにたどり着けないことや，ページに対するジオパースと空間インデックスを作成する処理コストが比較的高いという特徴がある. そのため，有用なページの効率的な収集戦略がきわめて重要な問題となる. 前節(1)の研究事例はこの問題に対する方針を与えているが，全国に分布するさまざまな空間コンテンツを均等に収集するため，特定の地域やテーマに関して十分な質・量のコンテンツを収集するには非常に長い時間がかかる．そのため，まずキーワード検索によって対象となるページを絞り込み，残ったページに対してジオパースを行うという方法がより現実的である. キーワード検索を用いて Q を解くアルゴリズム A1 は次のように書ける.

A1. キーワード検索による空間コンテンツ収集
(1) 業種 c に関連するキーワード k を用意する
(2) k を含むページ集合を選択する
(3) w ($w \in W_k$) から場所表記 l を抽出する
(4) l に対応する地点 p_0 と位置 p の距離 d を求める
(5) d が閾値よりも小さい w からなる集合 W'を求める

しかし A1 によって得られる W'は，単に k と l を含んでいるというだけで, 求める解答 W と等価ではない. たとえば，あるユーザが「目黒区駒場４丁目にあるラーメン屋」の情報を検索しようとして，「駒場４丁目」と「ラーメン」というキーワードを入力したとする. しかしページが「目黒区駒場４丁目」という住所と「ラーメン」というキーワードを含んでいたとしても，次のような誤検索が起こることが多い.

ケース 1 場所の記述 l とキーワード k が文脈上対応していない
「駒場４丁目にある日本民藝館」に遊びに行った帰りに「新宿３丁目の来来軒でラーメンを食べた」というように，場所の記述 l（=「駒場４丁目」）とキーワード k（=「ラーメン」）が対応していない場合.

ケース 2 キーワード k が業種 c を正しく表していない
「ラーメンのおいしい作り方」を説明したページで, ページを提供している料理好きな個人の勤務先が「駒場４丁目」であるというように，キーワード k(=「ラーメン」)が業種 c(=「ラーメン屋」)を表していない場合．このようなページも空間コンテンツの１つではあるが，ユーザの検索意図には正しく答えていない.

そこで，W'に含まれるページの中から検索意図に一致しているものだけを選択する手法を追加した A2 によって，問題 Q を解くことができる.

A2. キーワード検索による空間コンテンツの収集
[空間コンテンツ候補の収集]
(1)〜(5) A1 と同じ
[検索意図に適合した空間コンテンツの選択]
(6) W' に含まれるページ集合から，l と k が文脈上対応しているものを選択した集合を W''とする
(7) W''に含まれるページ集合から，k が c を表しているものを選択した集合を W とする

さて，A2 の(6)および(7)は自然言語文章の意味的な

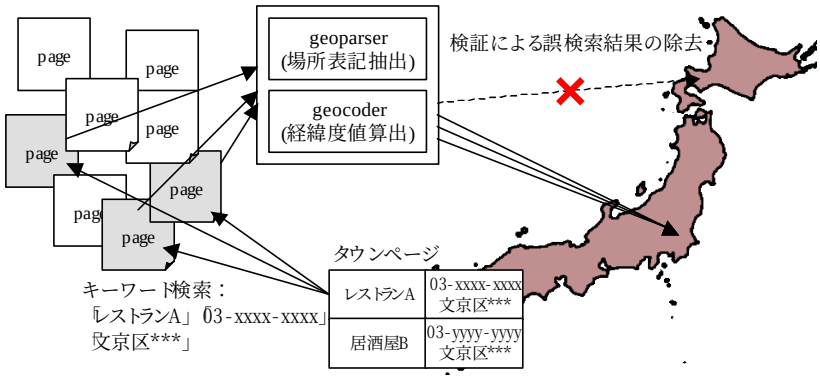

図 2 タウンページを用いた空間コンテンツ収集手法(提案手法)

理解を伴うため，計算機上で実現することが非常に難しい．そこで本研究では，店舗情報の辞書としてタウンページを用いることで，この問題をシンプルに解決した．おおまかには，検索キーワードとして「ラーメン」のような曖昧なキーワードを用いるかわりに，業種 c に含まれる店舗の名称や住所，電話番号を用いることで，より確実に業種 c と関連するページを収集する．また，タウンページの記載住所とページ w に含まれる場所表記を比較することで，l と k が対応しているかどうかを検証し，検索結果の精度を高める(図 2)．

4. 空間コンテンツ収集手法

本節では，タウンページ情報を用いて A2(1)から(5)に相当する部分について詳しく説明する．

4.1. キーワードの選択

まず検索キーワードを選択する手順について説明する．タウンページを「店舗名，住所，電話番号，業種」の辞書ととらえ，特定の業種に含まれる店舗名，住所，電話番号を検索キーワードとすれば，高い精度で空間コンテンツを収集することができる．また，店舗名をキーワードとすることで，上述のケース 2 のように業種 c とは無関係なページが誤って検索されるのを防ぐことが期待できる．

さて，実際にタウンページをみてみると，店舗名や住所に正式な表記が使用されており，これらの表記をそのまま検索キーワードとすると，部分的に省略されている文字列を含むページが検索できない．たとえば店舗名に含まれる「株式会社」などの表現は頻繁に省略されているし，住所の記述では都道府県名の省略や地番の表記が異なることがある．そのため，店舗名からは「株式会社」「有限会社」「＊＊法人」を取り除いたものをキーワード k_{s_name} とし，住所からは市町村・特別区名と町字名を抜き出した部分（例：目黒区駒場）をキーワード $k_{s\_address}$ として用いる．また，電話番号は市外局番が省略されることがあるので，市内番号部分をキーワード k_{s_tel} として利用する．

電話番号は一意性が高く検索キーワードとしては優れているが，電話番号が記載されていないページも多数存在すると考えられる．これまで実際にどの程度の店舗情報が Web 上に存在するのか調査した事例はなかったので，店舗情報の検索にどのようなキーワードを用いるのが有効かを調べるため，次のようなランクを定義して，実際に収集実験を行った．

定義：店舗情報を含む Web ページのランク

1. 漢字掲載名を品詞分解した文字列集合の一部が一致（例：「大塚 鍼灸 治療 院」）
2. 漢字掲載名（ただし「有限会社」「株式会社」「＊＊法人」の文字列を除去したもの）（例：「吉田歯科クリニツク」）が一致
3. 漢字掲載名と住所（市町村・特別区・町字）が一致（例：「よしむら歯科医院 文京区本郷」）
4. 電話番号が一致（例：「03-xxxx-xxxx」）
5. 漢字掲載名と電話番号が一致（例：「千石医院 03-xxxx-xxxx」）
6. 漢字掲載名，住所，電話番号がすべて一致

図 3 に，文京区のタウンページに記載されている全業種(16,272 件)を Google で検索し，上記のランクで分類した結果の分布を，全業種と特徴的な業種について示す．タウンページに記載されている全業種・全店舗の 8 割については，名称をキーワードとして検索すると何らかのページが見つかる(名称が同じ別の店舗の場合も含む)が，住所を含むものは 6 割程度である．対照的に，官公庁では 95%について名称と住所が一致するページが検索できる．また，医療機関では全体の半数について電話番号から何らかのページが検索できるが，飲食店では電話番号による検索は 35%程度，官公

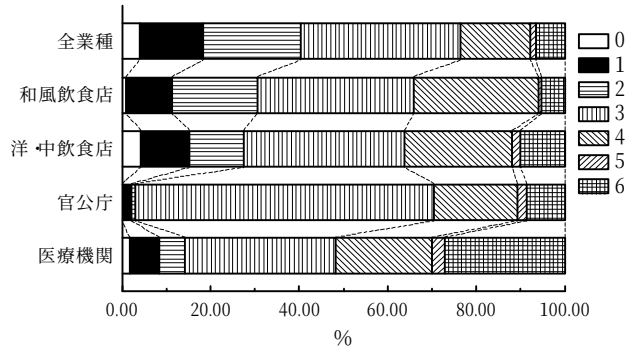

図 3 業種別検索レベルの割合

庁では 30%しか検索できない．

次に，いわゆるポータルサイトのカバー率を調べるため，図 4 に同実験によって得られたページのドメイン名を，業種ごとに上位 3 位まで示した．医療機関では「文京区のお医者さん」(www.bunkyo-med.or.jp)と「東京マイネット」(www.t-mynet.com)がそれぞれ全体の 50%弱を，官公庁では「文京区」(www.city.bunkyo.tokyo.jp)が 2/3 をカバーしている．飲食店では「ヤフーグルメ」(gourmet.yahoo.co.jp)が洋風・中華飲食店で 16%，和風飲食店では 13%弱をカバーしているのが最上位と，予想外にカバー率が低いことが分かった．言い換えれば，飲食店についてこれらのポータルサイトを利用して検索した場合，タウンページに記載されている店舗の 8 割以上は発見できないということであり，Web から直接ページを検索することに意義がある．また，電話番号だけでは検索できないページ（ランク 3 に相当）が飲食店全体で 40%近くあり，検索キーワードとして店舗名称と住所を用いる必要性が裏付けられた．

4.2. ページの検索と空間範囲検索

A2(2)のキーワード検索処理には Google を利用し，k_{s_name} だけを用いた検索結果上位 10 件，k_{s_name} と $k_{s\_address}$ を用いた検索結果上位 10 件，および k_{s_tel} を用いた検索結果上位 10 件の合計 30 件(1 店舗あたり)を収集する．

A2(3)および(4)の，場所表記を抽出し，対応する座標値を求める処理は，文献[9]で報告した「芭蕉」を用いて行う．地名辞書としては国土地理院発行の数値地図 25000(地名・公共施設)に含まれる 483,980 件の地名と，国土交通省によって整備されている全国街区レベル位置参照情報に含まれる 13,968,156 件の住所（番・地番レベル）を用いた．

A2(4)の距離計算は，厳密には平面直角座標系による計算を行うべきであるが，本研究で対象とする範囲はたかだか数 km 程度であり，次の式で十分近似できる．

業種名	URL1/URL2/URL3	率 (%)
医療機関	www.bunkyo-med.or.jp	47.65
	www.t-mynet.com	47.48
	www.zero-dr.jp	28.69
官公庁	www.city.bunkyo.tokyo.jp	66.91
	www.tokyo.japanpost.jp	25.18
	www.metro.tokyo.jp	10.07
教育	www.e-juq.net	15.16
	www.yubitoma.or.jp	13.41
	www.city.bunkyo.tokyo.jp	12.24
洋風・中華飲食店	gourmet.yahoo.co.jp	15.83
	homepage2.nifty.com	10.55
	www.geocities.co.jp	9.76
和風飲食店	gourmet.yahoo.co.jp	12.66
	www.geocities.co.jp	10.36
	www.sushi-all-japan.or.jp	7.29
総合工事及び測量・調査・設計	www.reform-net.com	20.78
	www.mokken.com	7.82
	www.mhlw.go.jp	7.28
旅行・旅館・ホテル	www.mytrip.net	12.84
	www.tokyo-hotel-ryokan.or.jp	10.09
	www.amenity.jp	6.42

図 4 検索結果に含まれるドメイン名上位 3 件と出現率

$$d^2 = (cos(y)\ (x_0 - x))^2 + (y_0 - y)^2 \qquad (1)$$

ただし x, y は p の経度・緯度，x_0, y_0 は l に対応する地点 p_0 の経度・緯度，d はこの 2 点間の直線距離である．

5. 空間コンテンツ検証手法

本節では，**A2**(6)および(7)について説明する．

5.1. ページのブロック分割

まず，**A2**(6)の「キーワード k と場所 l が文脈上対応している」という条件について考える．3 節の誤検索ケース 1 で示した例のように，キーワード k と場所 l が対応しない場合を検出するには，k と l が意味的に支配する文章領域を抽出する必要がある．このような領域抽出は一般の自然言語文章では困難であるが，HTML で記述されているページの場合，参考文献[7]で報告されているように HTML タグの構造を解析することで処理が可能である．ここではさらに，ページを l の支配領域であるブロックに分割する手法について説明する．

空間コンテンツを含むページには，2 つ以上の空間コンテンツを含んでいる場合がある．図 5 は本郷周辺の喫茶店を紹介したページであるが，1 ページ内に 10 件の喫茶店情報がレイアウトされている．このようなページから，k_{s_name} という名前で $k_{s\_address}$ という住所に存在する店舗をキーワードで検索すると，偶然 1 件目の店舗名称が k_{s_name} と一致し，3 件目の住所が $k_{s\_address}$ に一致するといった誤検索が起こることがある．これは希なケースに思われるが，実際には店舗名称として地名が用いられている場合（「団子坂」など）や一般的な語が用いられている場合（「山」など）があり，頻繁に発生する．

理想的には，店舗情報が 1 つずつ含まれるようにページをブロックに分割することができればよい．文献[7]では HTML 要素の種類（タグ名）によって領域を定

めているが，ここでは各ブロックに最低 1 つの場所表記が含まれるよう，HTML 要素を単位として分割する．

A3. ページのブロック分割

(1) HTML 要素をノードとする木構造を作成する

各ノードは，要素名称（<table>など），属性のリスト（width=”100%”など），タグで囲まれる文字列，および下位のノードのリストを属性として持つ．

(2) 場所表記数をカウントする

それぞれのノードに対し，文字列に含まれる場所表記の数をカウントする．子ノードに含まれる場所表記の数も合計する．

(3) 分割するノード木のレベルを決定

あるノードから見て，複数の子ノードが場所表記を含むとき，子ノード以下を別々のブロックとして分割する．ただし，ページ全体に場所表記が 1 つしかないなど，ブロックに分割できなかった場合には，ページ全体を 1 ブロックとする

図 5 複数の空間コンテンツを含むページ

図 6 は，典型的な表形式の店舗一覧から作成した木構造の例である．場所表記は表の 2 列目のセルに含まれているので，tr ノードはそれぞれ 1 つの場所表記を含む．table ノードから見たとき，複数の子ノード（= tr ノード）が場所表記を含んでいるので，各 tr ノード以下をブロックとして分割する．以上の処理により，表の各行をブロックとする分割を行うことができる．また，この手法ではブロックが同じ種類のタグで構成され同じレベルに並んでいれば，タグの種類に依存せずに利用できる．

5.2. キーワードの検証

A2(7)の「キーワード k が業種 c を表している」という条件は，本来は文章の意味的な理解が必要でありきわめて判定が困難であることから，ここでは「ブロック w から抽出されたキーワード集合$\{k_w\}$を，4.1 節で作成したキーワード集合$\{k_s\}$:=$\{k_{s_name}, k_{s_address}, k_{s_tel}\}$と比較し，同じ店舗 s を表していると認められる」ことで代替する．これは同時に A2(6)の条件である「キーワード k と場所 l が対応している」という条件の十分条件でもある．まず，店舗名称・住所・電話番号を比較する際の基準を示す．

A4. キーワードの比較基準

(1) 店舗名称の比較

ブロックの文章から店舗名称 k_{s_name} を文字列レベルで検索し，見つかれば名称が一致したとする．ただし，半角カナと全角カナ，漢数字とアラビア数字の違いは許容する．

(2) 住所の比較

ブロックの文章から「住所らしい」表記を抽出し，対応する地点 p_i(i = 1..n, n は見つかった表記の数)を算出する．タウンページに記載されている住所に対応する地点 p と p_i の距離 d_i を式(1)によって計算する．文字列として比較しないのは，市の名前が変わったなどの理由により同じ場所を表す住所が違う表記をされている可能性があるためである（「さいたま市」と「浦和市」など）．min(d_i)が 10m 以下の場合には住所が一致したと見なす．10m 以上であっ

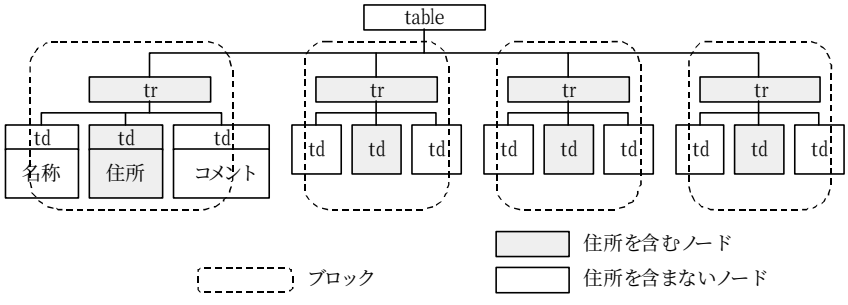

図 6 Web ページのブロック分割

ても「文京区湯島の来来軒」のように丁目・字で記述されている場合もあるので,100m 以下であれば住所一致の可能性ありとする.100m を超える場合には住所不一致とする.

(3) 電話番号の比較

電話番号 k_{s_tel} をブロックの文章から文字列レベルで検索する．全角数字や漢数字で表記されている場合には，あらかじめ半角アラビア数字に変換した上で処理を行う．市内番号部分が見つかった場合には一致の可能性あり，市外局番まで一致する文字列が見つかった場合には一致とする．

店舗の名称が一致していても同じ名前の別の店舗である可能性もある．だが，名称も住所も一致していれば,おそらく同じ店舗であると考えてよい.そこで，一致したキーワードの組み合わせにより，同じ店舗である確度 r を次のように定義する．

A5. ブロックに店舗 s の情報が含まれる確度 r

(1) 電話番号が一致した場合，$r = 4$

(2) 店舗名称と，電話番号の市内番号部分が一致した場合，$r = 4$

(3) 店舗名称と住所が一致した場合，$r = 3$

(4) 店舗名称が一致し，住所が 100m 以内の場合，$r = 1$

(5) 上記のいずれにも該当しない場合，　$r = 0$

5.3. 順位付け

以上の処理により，検索対象である店舗に関連するブロックをページから抽出し,列挙することができる.空間サーチエンジンとして利用する場合，抽出した結果を単に羅列するのではなく，有用な情報を含む順に並べて提示するためのスコアリングも検討するべきである．前節で定義した確度 r は，求める店舗の情報がページに含まれているかどうかを表しているので，スコアリングにも有効であると考えられる．しかしブロックが適切に分割できずに複数のコンテンツが含まれてしまっている場合や，店舗名称と電話番号以外は何の情報も含まれていない場合でも，電話番号さえ含まれていれば $r = 4$ となってしまうため，次のように修正する(ただしもともと $r < 2$ の場合には修正しない).

A5'. 確度 r の修正

(1) ブロック内に複数の場所表記が含まれている場合

5.1 で示した手法では，複数の店舗に関する情報を，タグを挟まずに書き綴った文章はブロックに分割することができないため，１つのブロック内に複数の場所表記が含まれている場合には，$r = 2$ とする.

(2) ブロックが，複数の店舗に検索結果として指されている場合

5.1 で示した手法では，場所表記を含まずに店舗情報を列挙してある場合にも分割できない．このようなブロックは複数の店舗から検索結果として指されるため，最後に検出可能である．この場合にも $r = 2$ とする.

(3) ブロックに含まれる文字数が閾値より少ない場合

ブロックに含まれる文字数は，そのブロックが持つ情報量をおおまかに近似していると考えられる．そこで，文字数が閾値（実装では 128 文字）より少ない場合には $r = 2$ とする.

6. 提案手法の評価と実装

6.1. 評価実験

提案手法の有効性を確認するため，収集したページと主観的な有効性の関係を調べる次の実験を行った.

実験 サンプリングによる有用性の判定

文京区タウンページに掲載されている飲食店からランダムに 1 件選択する．その店舗に関するページを提案手法によって収集し，そのうち 1 ページをランダムに選ぶ．選択したページを被験者に示し，その店舗に関する情報としての有用性を次の選択肢から選択させる(図 7).

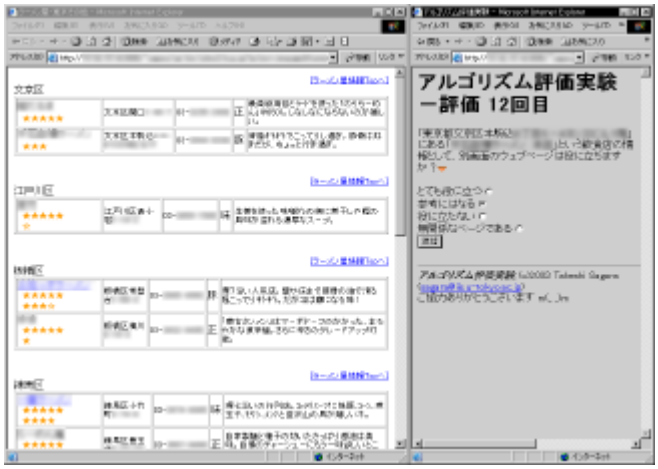

図 7 提案手法の有効性評価実験

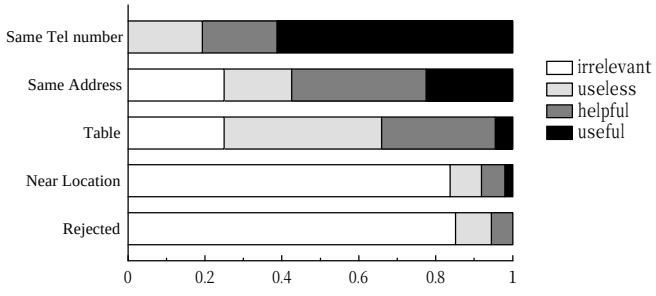

図 8 有効性評価実験の回答分布

図 9 空間サーチエンジンの実装例

(1) とても役に立つ
(2) 参考にはなる
(3) 役に立たない
(4) 無関係なページである

この結果と，A4(4)の判定基準とを比較したものを図 8 に示す（回答数 1,215 件）．*y* 軸は上から順に *r* = 4, 3, 2, 1, 0 となっている．空間コンテンツとして価値があるのは「とても役に立つ」(useful)，「参考にはなる」(helpful)と判定されたページなのでこれらの割合に着目すると，*r* = 4 の場合には 80%，*r* = 3 の場合には 57.5%，*r* = 2 の場合には 34%，それ以外では 7%程度である．このことから，この順にユーザに提示することで，より有用なページを優先的に示せることが分かる．

6.2. 実装例

図 9 は，提案手法を実装した空間サーチエンジンの一例である，Web アプリケーションの画面である．このアプリケーションでは，文京区のタウンページに含まれる飲食店（和風飲食店，洋風・中華飲食店，スナック・バー・酒場・喫茶店）1,986 件の情報を地図上で検索，閲覧することができる．各店舗の地図上の位置はタウンページに記載されている住所から算出できるため，ページ内に電話番号しか記載されていない場合でも，正しい位置にリンクすることができる．

また，一覧の部分には，ページ全体の文章を表示するのではなく，検索にマッチしたブロックに含まれる文章だけを表示することで，可読性を高めている．

7. おわりに

タウンページを辞書としてキーワード検索を行うことで，Web 上に存在する空間コンテンツを効率的に発見・収集・検証し，地図上にリンクする手法について述べた．従来手法ではリンクをたどりつつページを収集するため，特定のドメインを指定する以外にトピックや地域を限定することができなかった．そのため，必要な数のページを収集するのに数ヶ月かかることも珍しくない．一方，今回文京区の飲食店情報を収集するのに要した時間は，検証にかかる時間も含めて 24 時間程度であり，従来手法と比べきわめて短期間で情報を収集することができた．これは変化の速い Web ページに対しても十分実用的な収集能力である．

今後の課題としては，整備範囲を業種・地域ともに広げることと，実装したアプリケーションのインタフェース部分の改良が挙げられる．

謝辞　本研究を始めるきっかけとなった貴重なアイディアを頂きました貞広幸雄様，貴重な地図データを研究用にお貸し頂きました株式会社ゼンリン殿，研究を進める上で議論に参加して頂いた中山麻衣様，野秋浩三様に感謝いたします．

文　献

[1] Kevin S. McCurley, “Geospatial Mapping and Navigation of the Web”, ACM WWW10, 2001
[2] 横路誠司，高橋克巳，三浦伸幸，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
[3] 三浦信幸，横路誠司，高橋克巳，島健一，“GIS を用いた位置指向の WWW サーチエンジン～モーバイルインフォ 2 実験～”，地理情報システム学会講演論文集，Vol.7, pp.131-136，1998
[4] 大槻洋輔，佐藤理史，“地域情報ウェブディレクトリの自動編集”，情報処理学会論文誌，Vol.42, No.9, pp.2310-2318，2001
[5] 関根聡，“テキストからの情報抽出”，情報処理学会誌，Vol.40, No.4, pp.370-373，1999
[6] 竹元義美，福島俊一，山田洋志，“辞書およびパターンマッチルールの増強と品質強化に基づく日本語固有表現抽出”，情報処理学会論文誌，Vol.42, No.6, pp.1580-1591，2001
[7] 佐藤 理史，“ワールドワイドウェブを利用した住所探索”，情報処理学会論文誌，Vol.42，No.1, pp.59-67，2001
[8] 相良毅，有川正俊，坂内正夫，“分散位置参照サービス”，情報処理学会論文誌，Vol.42，No.12, pp.2928-2940, 2001
[9] 相良毅，有川正俊，坂内正夫，“ジオリファレンス情報を用いた空間情報抽出システム”，情報処理学会論文誌：データベース，Vol.41，No. SIG6(TOD7), pp.69-80, 2000