Web からの効率的な新規店舗の発見・登録支援手法

An Efficient Method to Support Finding and Registering New Shops from the Web

相良 毅♥　喜連川 優♥

Takeshi SAGARA　Masaru KITSUREGARA

Web から地理情報を抽出する手法には，クローリングによって得られた情報を，住所表記などを手がかりとして対応する場所に関連づける非登録型検索手法と，あらかじめリストを用意した店舗などの検索対象に関連づける登録型検索手法がある．非登録型手法は主題が異なるページを収集してしまうことがあり精度が低下する．登録型手法はより多くの情報を高い精度で収集できるが，リストに登録されていない対象に関する情報は収集できない．そこで，登録型検索手法により収集した Web ページを対象として非登録型検索手法を援用することにより，リストにない新規店舗を高い精度で検索できる手法を提案し，登録支援システムを開発した．

To extract geographical information from the Web, there are two typical approaches. The 1st one is preparing all geographical entities as a list, and crawled web pages will be linked to them by analyzing their content. The other one is retrieving web pages on demand with keywords given by the user, extract addresses from the pages to locate them to the ground. The 1st approach is more precise and able to acquire more information in general, however, no entities on the list can not be retrieved. Therefore, we have applied the 2nd approach to find new shops which are not on the list, from the collection of web pages retrieved by the 1st approach which contains many shop information in high probability. A prototype registration support system is also developed.

1. はじめに

特定のキーワードを含むページを一覧表示する既存のサーチエンジンでは，人物や店舗，家電製品のような固有の対象物（以下，便宜のため「エンティティ」と呼ぶ）に関する情報を検索しようとした場合，検索語に多義性が存在する（たとえば同姓同名・同名店舗・同名製品が存在する）場合には，キーワードだけで区別することは不可能である[1]．そのため，利用者が一覧に含まれる短い文章（スニペット）を見て判断する方法が一般的である．

一方，近年盛んになっているWebからの評価情報抽出手法や情報の信頼性に関する研究では，特定のエンティティに関する情報を多数収集することにより，そこから頻度の高い評価表現（「良い・悪い」など）を抽出したり，他の類似エンティティとの関係から情報源の信頼性を算出するといった処理が行われる[2]．これらの処理では入力情報の精度が出力結果を大きく左右するため，エンティティに関する情報だけを人手によらず高い精度で集める手法が必要である．そのため評価情報を抽出する情報源を特定の電子掲示板（BBS）などに限定することにより，そのページの構造からエンティティと確実に対応する情報が利用されることが多い[3]．特定の情報源に限定することは精度を高める上で有効であるが，情報の網羅性という点から見れば全Webページを対象とした方が優れている．そこで，まず全Webからキーワードを含むページを収集し，その中から検索キーワード以外の固有表現（Named Entity）を自動抽出することにより，キーワードに多義性が存在しても固有表現によって自動的に情報を分類するアプローチが考えられる[4]．

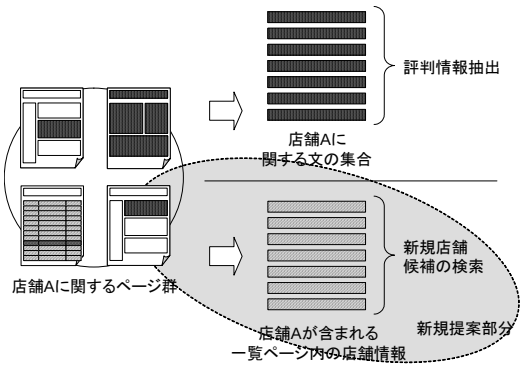

図 1 提案手法の概要
Fig 1. Proposed method overview

われわれは，実世界店舗を対象としてWeb から評判情報を収集し，店舗のランキング結果などと合わせてユーザに提示することにより，ユーザの意志決定を支援する「店舗情報検索システム」を開発している [5]．店舗のような地理的なエンティティを分類する固有表現としては住所や電話番号が抽出しやすくかつ識別能力も高いため，電話帳を辞書として利用している．電話帳は比較的網羅性が高く，入手しやすいという長所があるが，情報が時間とともに劣化するため常に更新しなければならない点と，電話帳に記載されていない店舗も多く存在する点が欠点である．そこで本稿では，電話帳に登録済みの店舗に関する情報を抽出するために収集したWeb ページのうち，一覧表などのタグ構造を手がかりに未登録の新規店舗候補の情報を抽出することにより，高い精度でかつ短時間に新規店舗を発見し，人手による登録を支援する効率的な手法を提案する（**図1**）．

以下，2 章で関連研究を，3 章で提案手法の詳細を示す．4 章で検索・登録支援システムの実装について説明し，5 章で実験と考察を行い，6 章でまとめる．

2. 関連研究

2.1 Web からの地理情報収集手法

Web から地理情報を収集するには，クローラでページを収集し，ページに含まれている地理的な場所を表す記述を抽出する手法が用いられる[6]．場所を表す記述は，住所のように位置を表すものと，ランドマーク（ビル名，施設名）のように地理的エンティティを表すものの2通りに分類することができる．

施設や店舗などの地理的エンティティを識別する研究と

♥ 正会員　東京大学 生産技術研究所
{sagara, kitsure}@tkl.iis.u-tokyo.ac.jp

して，大槻らは施設名をユーザが与えると，同名の施設が多数存在していてもその電話番号と住所の組を自動生成し，候補一覧を返す手法を開発した[7]．また，長屋らは，施設を想起させるキーワードをユーザが与えると，Google API を用いてページを収集し，ページ解析後に住所を抽出して地図上に位置を表示するシステムを開発した[8]．これらの手法ではエンティティを辞書に登録しておく必要がないため，長屋らの言葉を借りて「非登録型」地理情報検索手法と定義する．

非登録型検索手法は一般に，ユーザがキーワードを与えてから Web ページを収集し，住所や電話番号などの固有表現を抽出して地理的エンティティに関連づけるオンデマンドな Web 検索に用いられる．場所をキーとする Web ページ(またはその一部）の一覧が結果として得られるため，地図にして表示することもできる．非登録型検索手法ではユーザが与えたキーワードが適切であれば，どのような地理情報でも Web から収集し，地図として提示できるという点で優れている．その反面，クローリングや解析処理に時間が必要なこと，同一地点に複数の地理的エンティティが存在する場合に判別が難しいことから，収集した情報の適合率は比較的低く，収集できる件数も数百件が限界である．長屋らの実験によれば，特定のキーワードを含む地理情報を Web から収集した際，その情報がキーワードに適合する割合は，収集時間 30 秒の場合で 51.52～97.30%，60 秒の場合で 50.00～74.75%である（参考文献[8]中，表 2 で地名とキーワードを与えた 1,3,4,5 の 4 ケースより）．

一方，われわれの研究では，店舗の評判情報を収集するため，まずその店舗に関する情報だけを高い精度で網羅的に収集する必要があり，検索したい領域（地理的な範囲「新宿駅のそば」と業種「ラーメン店」）の店舗リストをあらかじめ準備している．これを「登録型」地理情報検索手法と定義する．登録型検索手法は一般に，連続的に Web 情報を収集し，各エンティティに情報をリンクさせて蓄積する．各エンティティに経緯度のような座標が与えられていれば，地図にして表示することもできる．収集した Web 情報がどのエンティティに関するものであるかを判別するために，住所や電話番号など各エンティティの持つ属性を利用することができるため，非登録型検索手法に比べて高い適合率で情報を取得できる．われわれの実験では，住所と電話番号を用いて店舗を識別することにより，内容の主題が検索目的と異なるもの（レストラン情報の検索に対し，レストランで開催される同窓会の案内ページを返すなど)を誤検索と見なしても約 84%，店舗の情報であれば正解と見なせば約 94%の適合率で店舗に関連する情報を収集することができた[9]．しかし，あらかじめエンティティの辞書を用意しなければならないこと，用意したエンティティに関する情報しか収集できないことが欠点であり，そのため検索の自由度は非登録型検索手法よりも劣る．

2.2 住所録の自動作成手法

住所録を人手で構築するには大きなコストがかかるため，Web ページのような情報源から住所録を自動構築する手法も研究されている．住所や電話番号は文字パターンや辞書を用いることで比較的容易に高い精度で抽出することが可能だが，店舗名のような固有名にはほぼ無数のバリエーションが存在するため，自動抽出が難しい．そこで村山らは，Web ページの HTML タグ構造の繰り返しパターンに着目し，固有名・電話番号・住所の３つ組を抽出する手法（以下「M 手法」）を開発している[10]．M 手法は，(1) はじめに小規模な３つ組辞書（電話帳）を用意し，(2) ２組以上の３つ組が現れる表形式の HTML を見つけ，(3) HTML タグ要素をノードとする DOM 木を構築，(4) 繰り返しパターンから固有名・電話番号・住所に対応するノードの位置を決定し，(5) 辞書にない新しい３つ組を取得する．新たに取得した３つ組を辞書に追加して処理を繰り返すことにより，さらに多くの３つ組を見つけることができる．

電話帳を辞書として利用する点や，表形式の Web ページを情報源として利用する点などが M 手法と提案手法に共通している．一方，M 手法は住所録を完全自動で構築することを目的としており，３つ組の各要素が独立した DOM 要素である HTML タグ構造を持つページだけを対象としているのに対し，提案手法では登録支援を目的とするため，もっとも自動化が困難な「店舗名の抽出」は人手で行う代わりに，より広範なタグ構造を持つページを対象とする（より多くの店舗が発見できる）点が異なる．

3. 提案手法

3.1 提案手法の目的と概要

適合率を高く保ったまま検索可能な店舗を拡大するために，非登録型検索手法を援用して新規店舗を検索し，店舗データへの登録を効率的に支援することを考える．既存の非登録型検索手法や住所録自動作成手法を新規店舗の検索に用いる際の最大の問題は，既存のキーワードベースのサーチエンジンでは店舗情報を含む Web ページを選択的に収集することが難しいことである．店舗情報を含まないページを収集してしまうと，余分な処理により時間を浪費したり，店舗ではない地理的エンティティを抽出して検索結果の適合率を低下させる．そこで，特定のキーワードを含むページを全 Web ページから検索する代わりに，登録店舗の情報を収集中に得られた Web ページ群から検索することにより，新規店舗候補の検索を効率よく行う手法を提案する．

たとえば既に登録されているラーメン店 A があるとすると，A の情報が含まれている Web ページは登録型検索手法により常時クローリングされ，多数のページが収集される．収集されたページの中には A が存在する地域の（ラーメン店以外を含む）飲食店リストや，全国の有名ラーメン店リストといった一覧ページが含まれていることが期待できる．基本的なアイディアは，これらの一覧ページを情報源として用いれば高い確率で店舗情報が含まれているため，高速かつ高精度に新規店舗候補が発見できるだろう，というものである．

表 1 店舗マスタデータベースの項目
Table 1. Items in shop master database

項目	内容
店舗名	店舗識別名称
住所	所在地（正規化済み）
電話番号	電話番号（正規化済み）
最終更新時刻	店舗情報を最後に更新した日時（秒）
有名度スコア	関連ページ数等から算出した得点

提案手法の説明のため，ここで Web から店舗情報を抽出する手法（発表済み）を簡単に説明する．まず店舗マスタデータベース（**表 1**）より，最終更新時刻が古く更新が必要な店舗の店舗名・住所・電話番号を１組取得する．この店舗名や住所の一部（町字名），電話番号の一部（市外局番部分を除いたもの）を検索語として，Web アーカイブより最大 300

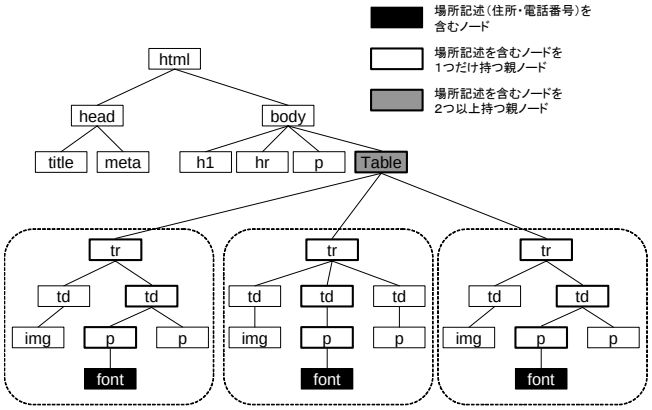

図 2 HTML 木の分割
Fig. 2. Dividing HTML Tag Tree

件の Web ページを取得する.収集した各 Web ページに対し，HTML タグ要素をノードとする木構造である HTML 木（一種の DOM 木）を構築し，住所または電話番号が 1 度ずつ含まれるように部分木に分割する（**図 2**）．各部分木に含まれるテキストセグメントのうち，名称や電話番号による検査をパスしたものを対象店舗に関する情報として取得する．

3.2 新規店舗候補の収集

提案する手法は，対象店舗の情報ではないために除去されていたテキストセグメントを新規店舗候補の情報源として収集し，のちにユーザが確認して登録する作業を行いやすい形にインデックスを生成して蓄積する．

アルゴリズム 1：新規店舗候補収集アルゴリズム

(1) 更新対象店舗である条件を満たさなかった部分木のテキストセグメント（部分木に含まれる住所と電話番号の情報を含む）を取得する
(2) 得られたテキストセグメントに含まれる電話番号が既に店舗マスタデータベースに登録済みであれば，新規店舗ではないので，そのテキストセグメントごと除去する
(3) テキストセグメントに含まれる住所，電話番号，およびテキストを新たなレコードとして新規店舗候補データベース（**表 2**）に登録し，検索インデックスを更新する

最終的に，アルゴリズム 1 によって得られる新規店舗候補のデータには，表 2 の項目が含まれる．

表 2 新規店舗候補データベースの項目
Table 2. Items in new shop candidates database

項目	内容
URL	元のページの URL
住所	抽出した住所（街区レベル）
経緯度	抽出した住所から得た経緯度
電話番号	抽出した電話番号
テキスト	テキストセグメント
更新日時	元のページのタイムスタンプ

3.3 新規店舗候補の検索と登録

新規店舗候補は，ユーザが特定の地理的範囲とキーワードを指定し，新規店舗候補データベースから該当するレコードを検索することにより提示される．提示された新規店舗候補を目視で確認し必要な修正を行った上で，店舗マスタデータベースに新たなレコードとして追加する．

アルゴリズム 2：検索・登録時アルゴリズム

(1) 新規店舗候補データベースから，ユーザが指定した地理的検索範囲内（表示されている範囲など）で，検索キーワード（「ラーメン」など）を含むレコードを検索する
(2) 該当するレコードを電話番号によりグループ化する
(3) 新規店舗候補のリストと，各候補の情報源となった Web ページから抜き出したテキストをユーザに提示する
(4) ユーザは提示されたテキストを参照しながら，必要な修正（店舗名の入力や住所・電話番号の確認）を行い，店舗マスタデータベースに登録する
(5) 登録された店舗と同じ電話番号を持つレコードを，新規店舗候補データベースより削除する

4. 検索・登録支援システムの実装

提案手法を用いた新規店舗候補の検索・登録支援システムの実装について説明する．新規店舗候補の収集処理（アルゴリズム 1）は従来の店舗情報抽出処理と並列してサーバ上で継続的に行われ,候補データを自動抽出して RDBMS 上に格納する．**表 3** に，アルゴリズム 1 により実際に収集した店舗候補データベースの大きさを示す．ただし，表中の「店舗数」とは，電話番号でグループ化した際のグループ数である．なお，このデータベースには，アルゴリズム 1(2)の処理により，店舗マスタデータベースに登録済み店舗と電話番号が重複するレコードは含まれていない．

表 3 新規店舗候補データベースのサイズ
Table 3. Size of new shop candidates database

データ種類	レコード数	店舗数
全データ	4,683,368	970,359
「ラーメン」を含む	74,848	10,100
「そば」を含む	45,445	8,959
「居酒屋」を含む	52,516	18,426

検索インタフェースは登録型店舗検索システムを拡張した Web アプリケーションとなっている（**図 3**）．ユーザが「新規店舗も検索」チェックボックスにチェックした場合，登録済み店舗に加え，アルゴリズム 2 を用いて新規店舗候補を検索し一覧表示するとともに(図中,左下のリスト C～E)，地図上の対応する位置にアイコンを置く．地図上の各アイコンは経緯度をキーとして新規店舗候補データベースの各レコードにリンクされており，アイコンをクリックするとその地点に存在する新規店舗候補の情報を含むページの URL とテキストセグメントの一覧が表示され，自動抽出された住所と電話番号が入力ボックスに入る（図右側）．登録をおこなうユーザは，テキストセグメントからカット・ペーストするなどして店舗名を入力し，住所と電話番号を確認する．登録は即時行われ，登録された店舗は電話帳から登録された他の店舗と同様に，クローリングのスケジュールに追加される．一定時間が経過し,十分な数の Web ページが収集されれば，評判情報などを抽出することができる．

5. 実験・考察

提案手法の有効性を検証するため，次の実験を行った．

実験 検索対象の違いによる性能の比較

提案手法は，あるキーワードに対して非登録型検索手法（アルゴリズム 2）を用いて店舗候補を検索する際，検索対象を全 Web ページとするよりも，アルゴリズム 1 によって用意した新規店舗候補データとした場合の方が適合率が高ければ，有効であると考えられる．そこで，３種のキーワー

図 3 実装した新規店舗検索・登録支援システムの画面例
Fig.3. An Implementation of the new shop retrieving, registering support system

ドに対して検索を行い，結果を比較した（表 4）．

表 4 検索適合率の比較結果
Table 4. Retrieval Precision Rates

検索語	全 Web	提案手法
ラーメン	0.725	**0.955**
そば	0.687	**0.940**
居酒屋	0.924	**0.975**

正解かどうかの判定は人手で行い，キーワードから一般的に予想される店舗情報であれば正解とした．つまり「そば」で検索した場合，「日本蕎麦」や「中華そば」は正解とするが，「そば粉（製粉所）」や「そば打ち教室」のような結果は不正解とする．全 Web ページからの検索は事実上不可能なため，検索語に「住所」を加えて Google で上位 10 ページを検索し，その下位ページから住所が含まれているものを母数として，正解であるものの割合を適合率とする．1 ページ中に複数の店舗が含まれている場合は検索語が含まれているセグメントのみを対象とする．提案手法を用いた場合，地域を限定しない場合には多くの結果が返されるため，表 3 に示した新規店舗候補データベースよりランダムに 200 件を選択して確認した．

実験では 3 種類の検索語について適合率を検証し，いずれの場合も提案手法が全 Web を対象とする場合よりも高い精度で店舗候補を検索できることが確認できた．これは，提案手法では情報の抽出が容易な一覧表形式に限定していることと，店舗情報と共起する地理情報を検索対象としているためである．その結果，どのキーワードに対しても 9 割以上と高い精度で店舗情報が得られる．

6. おわりに

店舗マスタデータベースに登録済みの店舗に関する情報を抽出するために収集した Web ページを情報源とすることにより，未登録の新規店舗情報を高い精度で抽出し，人手による登録を支援する効率的な手法を提案した．また，取得した新規店舗候補が店舗である精度（適合率）が高いことを実験により示し，提案手法の有効性を確認した．

提案手法の問題としては，最初の検索キーワードをユーザが考える必要があるため，適切なキーワードが与えられないために登録されないままになってしまう店舗が存在する可能性があることが挙げられる．新規店舗候補データベースから何らかの方法によりマイニングを行い，システム側からユーザに新規店舗を推薦するような手法を開発し，新規店舗を漏れなく見つけることが今後の課題である．また，店舗の中には，ブログなど個別の Web ページ上には紹介されても，一覧表形式のページにはなかなか情報が掲載されないケースが存在する．これらの情報も Web から抽出する手法の開発が必要である．

[文献]

[1] Ron Bekkerman, Andrew McCallum: “Disambiguating Web Appearances of People in a Social Network”, WWW2005, pp. 463－470 (2005)

[2] Kushal Dave, Steve Lawrence, David M. Pennock: “Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews”, WWW2003, pp. 519－528 (2003)

[3] 立石 健二，石黒 義英，福島 俊一:“インターネットからの評判情報検索”，情報処理学会研究報告，NL-144-11, pp. 75－82 (2001)

[4] 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)

[5] 相良 毅，牧野 俊朗，川口 修一，小澤 英昭，喜連川 優:“住所情報を用いた店舗名称のクリーニング手法”，データ工学ワークショップ 2006, 2C－o1 (2006)

[6] 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)

[7] 大槻洋輔，佐藤理史: “地域情報 Web 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)

[8] 長屋 務，森本 泰貴，藤本 典幸，出原 博，萩原 兼一: “Google Maps API を応用したロボット型施設検索システムの試作”，データ工学ワークショップ 2006，5B－i6 (2006)

[9] 相良 毅，喜連川 優: “日常生活をより豊かにする Web マイニング”，第一回横幹連合コンファレンス，E1－32 (2005)

[10] 村山 紀文，南野 朋之，奥村 学: “メタデータ付与のための住所録自動生成”，情報処理学会研究会報告－自然言語処理，Vol.2004，No.73，pp. 41－47 (2004)

相良 毅 Takeshi SAGARA
東京大学生産技術研究所戦略情報融合国際研究センター助手．1995 年東京大学大学院工学系研究科修了．Web マイニング，地理情報システムの研究に従事．情報処理学会，日本データベース学会，地理情報システム学会会員．

喜連川 優 Masaru KITSUREGAWA
東京大学生産技術研究所教授．同所戦略情報融合国際研究センター長．1983 年東京大学大学院工学系研究科情報工学専攻博士課程修了．工学博士．データベース工学，並列処理，Web マイニングに関する研究に従事．本会理事，情報処理学会フェロー，SNIA-Japan 顧問，ACM SIGMOD Japan Chapter Chair (H11-H14)，電子情報通信学会データ工学研究専門委員会委員長(H9,10)．VLDB Trustee, IEEE TKDE Assoc. Editor, IEEE ICDE, PAKDD, WAIM Steering Comm. Member.