Webからの効率的な新規店舗の発見・登録支援手法

相良 毅[†] 喜連川 優[†]

Webから地理情報を抽出する手法の1つに，あらかじめ検索対象のリストを作成し，クローリングによって得られた情報を各検索対象に関連づける登録型検索手法がある．登録型検索手法は，リストを用意せずにオンデマンドに検索を行う非登録型検索手法に比べ，より多くの情報を高い精度で収集できるという長所があり，評判情報抽出など情報の精度を必要とする処理には適しているが，リストに登録されていない対象に関する情報を収集することができないという欠点がある．そこで，登録型検索手法により収集された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し登録できる手法を提案し，登録支援システムを開発した．

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

Takeshi Sagara[†] and Masaru Kitsuregawa[†]

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, so the approach is suitable for reputation / opinion extraction, however, no entities on the list can not be retrieved by the approach. Therefore, we have applied the 2nd approach to find new shops which are not on the list, from the web pages retrieved by the 1st approach. Since the web pages retrieved by the 1st approach contain many shop information in high probability, the proposed method can extract new shops efficiently. 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，Web上の情報が増大するにつれ，検索目的とは異なるページが多数表示されてしまうという問題がある．特に，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・同名製品など）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）を見て判断することにより，目的の情報を選別するのが一般的である．

一方，近年さかんになっている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したり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高い精度で集める手法が必要であるが，上述した多義性の問題のために，全Webページを対象とした場合には高い精度で情報を収集するのは困難である．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うえで有効であるが，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，そこ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

† 東京大学生産技術研究所
Institute of Industrial Science, The University of Tokyo

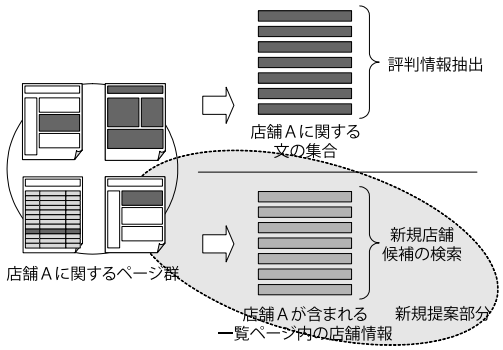

図 1 提案手法の概要
Fig. 1 Proposed method overview.

プローチが考えられる[4]．たとえば人物では，所属する企業名のような固有表現が抽出できれば，同姓同名の人物であっても別人であることが分かる．

我々は，実世界店舗（いわゆるオンラインショップではない店舗）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．そのため，住所・電話番号の辞書として電話帳を利用している．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．しかし，電話帳に含まれる情報は時間とともに劣化するためつね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．そこで本論文では，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 Web から収集したページから，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．

提案手法の概要を**図 1** に示す．これまでに開発した店舗情報検索システムは，まず電話帳に記載されている店舗の名称，住所，電話番号などを検索キーワードとして，Web から当該店舗の情報を含むページをクローリングする．次に，収集したページから当該店舗に関する部分を抜き出し，索引語や評価表現を手がかりに，評判情報を抽出する．一方，今回新規に提案する手法では，これまで利用していなかった「当該店舗の情報ではない部分」を活用する．直感的には，店舗が含まれている一覧表形式の Web ページを探し，そこから電話帳に記載されていない店舗情報と思われる表記を抽出することにより，新規店舗の候補とする．

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 章で実験結果を示す．6 章で考察を行い，7 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．住所を抽出する場合，住所表記の辞書と経緯度に変換するジオコーディングエンジンが必要で，その地点に複数のエンティティが存在する場合（雑居ビルに複数の店舗が入居している，など）にはどのエンティティに関する情報かの判別が難しいが，幅広い地理情報を収集することができる．一方ランドマークを抽出する場合，ランドマークの辞書が必要で，辞書に登録されているランドマークに関する地理情報しか収集できないが，個々のエンティティに直接情報をリンクすることができる．

2.1.1 非登録型地理情報検索手法

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オンデマンドな Web 検索に用いられる．つまり，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．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．このように，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．

一方で，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）．この適合率は，ユーザが目視により情報を取捨選択する場合には有用な値であるといえる．

2.1.2 登録型地理情報検索手法

我々の研究では，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．検索時には指定されたエンティティにリンクされている情報を返す．主キーはエンティティであるが，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．我々の実験では，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]．

登録型検索手法の欠点は，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である．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

地理的エンティティの辞書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，

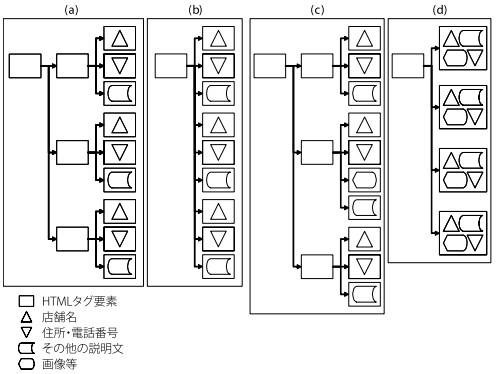

図 2 HTML 部分木の種類
Fig. 2 Types of HTML tag subtrees.

Web ページの HTML タグ構造の繰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 Web ページを見つけ，(3) HTML タグ要素をノードとする DOM 木を構築，(4) 繰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，最も自動化が困難な「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．違いを明確にするため，各手法で抽出可能な HTML タグ構造パターンを示す（**図 2**）．なお提案手法の詳細は 3 章で説明する．

- (a) 店舗名，住所，電話番号，説明文などが規則正しく階層的に構造化されたパターンで，主に table タグを用いた表に現れる．この場合には M 手法でも提案手法でも抽出できる．
- (b) 店舗名，住所，電話番号が店舗ごとに階層化されることなくフラットに並んだパターンで，主に br タグを用いて情報を羅列したページに見られる．各店舗の情報が部分木に分割できないため，M 手法では抽出できるが，提案手法では抽出できない．

- (c) (a) と同様 table タグを用いたページに見られるが，一部の構造に img タグによる画像などが含まれているなど，例外的なパターンが含まれている．この場合は提案手法では抽出できるが，M 手法では抽出できない．
- (d) 近年ブログなどによく現れるパターンで，各店舗の情報はエントリごとに分割されているものの，店舗名・住所・電話番号がタグで区切られずに文章中に記述されている．提案手法では抽出できるが，M 手法では抽出できない．

3. 提 案 手 法

3.1 提案手法の目的と概要

我々の店舗情報検索システムでは，評判情報の収集などのために適合率を優先させ，登録型検索手法を用いている．しかし，登録されていない新規店舗が検索できないこと，新規店舗を登録するためのコストが大きいことが問題となっていた．そこで，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たりする．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良く行う手法を提案する．

たとえばすで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し，多数のページを収集できる．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，高速かつ高い精度で新規店舗候補が発見できるだろう，というものである．

3.2 （参考）Web からの店舗情報抽出手法

提案手法の説明のため，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．

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

アルゴリズム 0：Web からの店舗情報抽出手法

(1) **更新対象店舗の決定** 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（住所は地名データベース☆を，電話番号は市外局番一覧情報☆☆を用いてあらかじめ正規化済み）．

(2) **関連 Web ページの取得** 店舗名，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 件の Web ページ URL を取得する．

(3) **粗い検査** 得られた Web ページのうち，i) 店舗名および住所の一部を含む，ii) 店舗名および電話番号の一部を含む，iii) 住所の一部および電話番号の一部を含む，iv) 市外局番を含む完全な電話番号を含む，のいずれかの条件を満たすものを残し，それ以外のページは対象とする店舗の情報が含まれていないと判断して除外する．

(4) **テキスト分割** 残った各 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 3**）．

(5) **ページ種別の判定** 部分木の数 n_{sub}，部分木に含まれるテキストセグメントの最大長 l_{max} に対し，$n_{sub} = 1$ の場合は店舗詳細ページ，$1 < n_{sub} \leq \theta_1$ かつ $l_{max} > \theta_2$（θ_1，θ_2 は任意の閾値，たとえば 10 と 256）の場合は店舗概要ページ，それ以外の場合（$n_{sub} > \theta_1$，または $n_{sub} > 1$ かつ $l_{max} \leq \theta_2$）は店舗一覧表ページとする．

(6) **詳細な検査** 各部分木に含まれるテキストセグ

☆ 街区レベル位置参照情報ダウンロードサービス，国土交通省，http://nlftp.mlit.go.jp/isj/

☆☆ 総務省の情報通信政策に関するポータルサイト，総務省，http://www.soumu.go.jp/joho_tsusin/top/tel_number/shigai_list.html

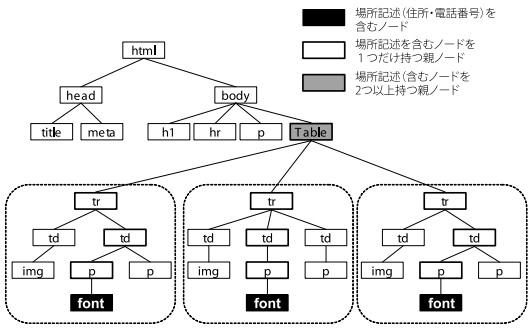

図 3 HTML 木の分割
Fig. 3 Dividing HTML tree.

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

メントに対し，上記 (3) で示した基準を満たすものを対象店舗に関する情報として取得する．

3.3 新規店舗候補の収集

提案する手法は，アルゴリズム 0 の (6) において，基準を満たさ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，あとでユーザが目視で確認しながら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) アルゴリズム 0，(6) の処理で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すで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．

3.4 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．そのユーザが登録権限を持つ場合，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うえで，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．

新規店舗候補検索時のアルゴリズムは既存の非登録型検索手法とほぼ同じであるが，検索対象とする Web ページが全 Web ではなく，アルゴリズム 1 によって作成した新規店舗候補データベースなので高速である点だけが異なっている．

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（地図ウィンドウに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) 該当する新規店舗候補レコードを，住所と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) で人手による確認が行われるため，店舗マスタデータベースに登録されるデータは十分に高い精度であることが期待できる．

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）は店舗情報抽出処理（アルゴリズム 0）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．ユーザが地理的な検索範囲を座標値で指定して新規店舗候補を検索する場合には経緯度を利用するため，住所を経緯度に変換して空間インデックスを作成する．キーワードを指定して新規店舗を検索する場合にはテキストから全文検索を行うため，単語に分解してフルテキストインデックスを作成する．また，新たな店舗が店舗マスタデータベースに登録された場合，同じ電話番号を持つ候補を新規店舗候補データベースから速やかに削除するため，電話番号にはハッシュインデックスを作成する．

表 3 に，アルゴリズム 0 により収集した約 516 万

図 4 実装した新規店舗検索・登録支援システムの画面例
Fig. 4 An Implementation of the new shop retrieving, registering support system.

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	5,019,096	1,069,316
「ラーメン」を含む	70,190	49,500
「居酒屋」を含む	56,516	29,324
「ランチ」を含む	42,984	28,456
「そば」を含む	51,662	34,362
「ワイン」を含む	14,707	9,600
「日本酒」を含む	6,287	4,541

Web ページから，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．1 ページ中に含まれる店舗情報の数はさまざまなので，ページ数とレコード数は一致しない．また，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，店舗マスター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．

一方，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 4**）．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．なお図で E が 2 度表示されているのは，同じ住所に異なる電話番号を持つ新規店舗候補が存在するためである．

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）．登録を行うユーザは，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．また，不適切な登録を防ぐため，現在はユーザ名とパスワードの入力が必要となっている．

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

提案手法の有効性を検証するため，以下の 2 つの実験を行った．

実験 1. 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

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

検索語	全 Web	D1	D2
ラーメン	0.725	**0.960**	0.905
居酒屋	0.924	**0.970**	0.970
ランチ	0.885	**0.965**	0.905
そば	0.687	**0.930**	0.815
ワイン	0.514	**0.955**	0.870
日本酒	0.440	**0.930**	0.825

る．そこで，6 種のキーワードに対して検索を行い，結果を比較した（**表 4**）．

なお，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「東京」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．また，アルゴリズム 1 (2) において対象としている登録店舗を含むページだけを用いたデータセットを D1，登録店舗を含まないページも用いたデータセットを D2 とする．D1，D2 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 2. 店舗データ登録に要する時間の比較

提案手法により新規店舗を登録する時間が短縮できることを，実際に店舗データを入力する時間を測定することにより検証した．

ケース A では，街頭で配布されている無料タウン情報誌☆から，店舗データベースに登録されていない店舗を 20 件見つけて名称だけのリストを作成しておき，この 20 件の情報を Web から検索し，住所と電話番号を登録するのに要する時間を測定した．

ケース B では，A と同じ 20 件のリストをもとに，提案手法で店舗名称を検索キーワードとして利用し，住所と電話番号を登録するのに要する時間を測定した．

ケース C では，A と同じ地域の飲食店で店舗データベースに登録されていない任意の 20 件を，さまざまなポータルサイトを用いて検索し，名称と住所，電話番号を登録するのに要する時間を測定した．利用するポータルサイトは被験者が自由に選択した．

☆ hot pepper，株式会社リクルート

表 5 新規店舗登録に要する時間（20 店舗）
Table 5 Time required for registering 20 new shops.

ケース	平均所要時間（秒）
A（店舗名リスト＋ Web 検索）	1,626.4
B（店舗名リスト＋提案手法）	482.1
C（Web ポータルサイト）	2448.2
D（提案手法）	331.4

ケース D では，A と同じ地域の飲食店で店舗データベースに登録されていない任意の 20 件を，提案手法を用いて検索し，名称と住所，電話番号を登録するのに要する時間を測定した．検索キーワードは被験者が任意に設定した．

ただし，使用したタウン情報誌は Web サイト上でも店舗情報を提供しているため，ケース A および C ではこのサイトを利用せずに検索を行った．また，このサイトは住所を含む一覧表形式のページを持たないため（店舗リストから任意の 1 件を選択し，詳細ページに移行すれば，各店舗の住所や電話番号が記載されている），提案手法では検索されない．そのためケースケース B，D ではこのサイトの情報は利用されていない．

実験の結果を**表 5** に示す．なお，ケース A，B は被験者 2 名により 1 回ずつ行った平均値，C，D は被験者 1 名により 3 回行った平均値である．

6. 考　　察

6.1 実験 1 に関する考察

実験 1 では 6 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．

検索語によっては，特に全 Web を対象とした場合に大きな差が生じた.「そば」の場合の適合率が低いのは,「駅のそば（近く）」のように「蕎麦」以外の意味で用いられることが多いためである.「居酒屋」は店舗情報以外で触れられることが少なく，全 Web を対象とした場合にも高い適合率で検索することができる.「ワイン」「日本酒」のようにそれ自体が検索対象となる検索語を用いた場合には，飲食店ではなく販売店の情報が多数検索されてしまうため，組み合わせる地域名（実験では「東京」）によっては適合率が大きく低下する．一方，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た結果，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．

次に D1 と D2 の適合率の違いについて考察する．D2 には，店舗マスタデータベースに登録されている店舗が 1 件も記載されていないページが含まれており，検索語から想定される業種とは無関係な住所一覧が検索されることがあるため，適合率を低下させる原因となっている．たとえば「そば」の場合「そば打ち教室」などが検索され,「ワイン」や「日本酒」の場合には販売店が検索されることがある．実験の結果から，登録済み店舗情報を含む一覧表のページに限定することにより，これらの誤検索が発生する確率が低下し，より高い適合率を得ることが可能であることが分かる．

6.2 実験 2 に関する考察

実験 2 では，あらかじめ新規店舗の名称リストを用意したケース A，B で 4 倍，用意していないケース C，D で 10 倍程度，登録に要する時間が短縮されている．後者の場合に特に効果が高いのは，ケース C では Web から検索した店舗候補がすでに登録されているかどうか確認するという冗長な処理に時間がかかるためである．また，店舗の住所，電話番号は自動的に抽出されているため，ユーザが行う処理は，(1) 提示されるテキストから店舗名を抜き出して入力し，(2) 電話番号，住所が正しいことをテキストから読み取る，という 2 つの処理だけでよい．

プロトタイプシステムを用いた実験では，登録済み店舗の除去や住所・電話番号の自動抽出が正しく機能することにより，登録時間が短縮されることが確認できた．よって，提案手法は新規店舗の検索・登録支援を行ううえで有効であるといえる．

ただし，実験 2 は被験者の Web 検索技術レベルや対象地域によって大きな差が生じること，実験に長時間を要するため十分な回数を試行できなかったことから，本実験の結果だけからは処理に要する時間が一般に 4〜10 分の 1 に短縮できるとはいえない．

7. 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，まだデータベースに登録されていない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，(1) 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，(2) 取得した新規店舗候補の情報を用いることで登録に要する時間が短縮されることを実験により示し，提案手法の有効性を確認した．

一方，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あげられる．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．このような場合には提案手法では新規店舗として検出することができない．さらに，提案手法では新規店舗は見つけることができるが，閉店または移転によって存在しなくなった店舗を見つけることができない．店舗データベースをより最新に近い状態で維持するためには，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．

謝辞 本論文を改善するにあたり，DBWeb2006 にて参加者の方々よりいただいたご質問・コメント，および査読者の方々の適切なご意見を参考にさせていただいた．ここに記して謝意を表す．また，本研究の一部は，文部科学省特定領域研究「情報爆発時代におけるサイバー空間情報定量評価基盤の構築」によるものである．

参 考 文 献

1) Bekkerman, R. and McCallum, A.: Disambiguating Web Appearances of People in a Social Network, *WWW2005*, pp.463–470 (2005).
2) Dave, K., Lawrence, S. and Pennock, D.M.: 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 1 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).

(平成 18 年 9 月 15 日受付)
(平成 19 年 2 月 27 日採録)

(担当編集委員 石川 博，有次 正義，片山 薫，
木俵 豊，中島 伸介)

相良 毅（正会員）

平成 5 年埼玉大学工学部情報工学科卒業．平成 7 年東京大学大学院工学系研究科情報工学専攻修了．平成 10 年東京大学空間情報科学研究センター助手．平成 15 年より同大生産技術研究所戦略情報融合国際研究センター助手．博士（工学）．空間情報科学，データ構造，Web 情報マイニングに関する研究に従事．本会データベースシステム研究会委員，電子情報通信学会データ工学研究専門委員会委員．

喜連川 優（フェロー）

昭和 53 年東京大学工学部電子工学科卒業．昭和 58 年同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．同年同大学生産技術研究所第三部講師．現在，同教授．平成 15 年より同所戦略情報融合国際研究センター長．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．平成 9 年，10 年電子情報通信学会データ工学研究専門委員会委員長，平成 11〜14 年 ACM SIGMOD Japan Chapter Chair，平成 14 年，15 年本会理事，平成 15 年本会フェロー．日本データベース学会理事，SNIA-J 顧問．IEEE TCDE Asian Coordinator，ACM SIGMOD Advisory Board Member.