# 住所情報を用いた店舗名称のクリーニング手法
## — Web からの店舗情報収集精度向上のために —

相良 毅[†] 牧野 俊朗[‡] 川口 修一[‡] 小澤 英昭[‡] 喜連川 優[†]

[†]東京大学生産技術研究所 〒 153-8505 東京都目黒区駒場 4-6-1 Ee-505
[‡]NTT レゾナント株式会社 〒 100-0004 東京都千代田区大手町 1-6-1 大手町ビルヂング 3F

E-mail: [†]{sagara, kitsure}@tkl.iis.u-tokyo.ac.jp [‡]{makino, shuichi, h.ozawa}@nttr.co.jp

**あらまし** Web からの店舗情報収集を行う際には，収集した Web ページがある店舗に関連することを確認するため，店舗名称を識別語として利用する必要がある．しかし，店舗データベースに登録されている店舗名称には，支店名など Web ページには記載されていない可能性のある語（不要語）が含まれているため，収集したページを正しく関連づけられないという問題がある．また，不要語にはビル名を用いた支店名など多くのバリエーションがあり，不要語辞書を整備して除去することは難しい．そこで，店舗データベースに含まれる住所の情報や，同じ住所に存在する複数の店舗名称を用いることにより，店舗名称をクリーニングする手法を開発した．実験によると，提案手法のクリーニング正解率は 91.6% と実用的な性能を示した．さらに，提案手法を導入することにより，従来より約 5% 多くの Web ページを収集できることを確認した．

**キーワード** 地域情報検索，住所情報，クリーニング，サーチエンジン

## 1. はじめに

近年，地域のイベントや，レストランなどの店舗情報，観光案内情報などエリアに特化した「地域情報検索」と呼ばれる Web サーチエンジンが多数登場し，注目されている [1][2][3][4][5]．地域情報検索では，通常のサーチエンジンの機能に加え，地理的な場所情報をキーとして検索する機能を実現するため，Web ページを実世界上の場所に関連づける処理が必要となる．この処理は手動で行われていることも多いが，自動で行う場合には，Web ページ内の文章に含まれる地理的な情報を抽出するジオパース (geo-parse) 処理が用いられる．ジオパース処理は大きく 2 つに分けられ，1 つは (**g-1**) ランドマーク名や電話番号といった識別語を抽出して実世界の「対象物」に関連づける方法，もう 1 つは (**g-2**) 地名や住所を抽出して実世界の「場所」に関連づける方法である [6] (**Fig.1**)．**g-1** の場合，対象物のデータベースを整備する必要があるが，高い精度で関連づけを行うことができる [7]．**g-2** の場合には対象物のデータベースは不要だが，その場所に存在する何について記述しているのか（地域の概要，あるいはそこに存在する施設に関する情報，もしくは施設内の店舗に関する情報）が分からないといった問題がある．

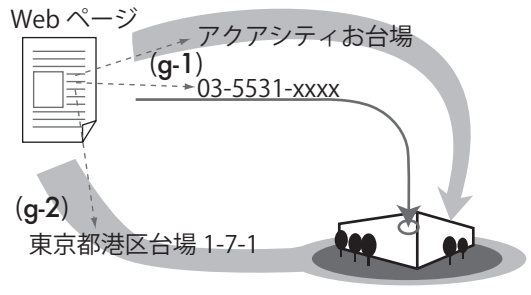


**Fig. 1 ジオパース手法による実世界への関連づけ**

地域情報検索の中でも，店舗情報は検索需要が高く，イベントなどに比べて対象となる件数が桁違いに多いため，自動的な処理によるメリットが大きい．また，既存の電話帳を店舗データベースとして利用でき，高い精度でジオパース処理が可能である．そこでわれわれは，実世界に存在する店舗の情報を Web から収集する手法を開発し [8]，店舗情報検索と呼んでいる [9]．

店舗情報検索では，Web ページを各店舗に関連づける必要があるため，ジオパース処理として主に **g-1** の識別語を用いる手法を利用している．その際に識別語の 1 つとして店舗名称を用いるが，店舗データベースに登録されている店舗名称には支店名などジオパース処理を行う上で精度を低下させるノイズとなる文字列（以下，不要語と呼ぶ）が多数含まれている．これらの不要語は，ビル名を用いた支店名など多くのバリエーションがあり，網羅的な辞書を整備して除去することが困難なため，効率的なクリーニング手法の開発が技術的課題となっていた．そこで，店舗データベースに登録されている店舗の住所の情報や，同じ住所に存在する複数の店舗名称を用いることにより，大規模な静的辞書を整備せずに，高い精度で店舗名称をクリーニングする手法を開発した．

開発した手法の概要は以下の通りである（**Fig. 2** 参照）．まず不要語には，普遍的に用いられるもの（「有限会社」など）と，場所に特有なもの（「六本木ヒルズ店」など）がある．前者は数も少なく静的なので，小さな (**α-1**) 不要語辞書を構築することで抽出できる．一方，

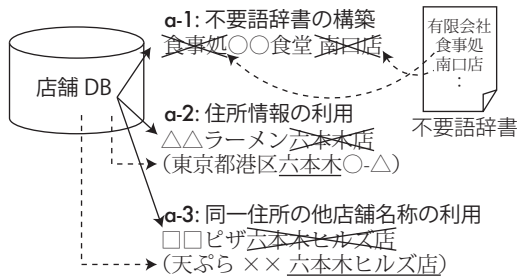


Fig. 2 店舗名称のクリーニング手法 ( 概要 )

後者は場所の名称（地名，駅名，ビル名など）にあわせて変化するため，網羅的な辞書を構築して維持することは困難である．そこで，(α-2) 店舗データベースに登録されている住所に含まれる文字列を利用する．たとえば住所が「東京都港区六本木○ - △」であれば，「六本木店」はおそらく支店名であって店舗特有の名称ではないことが推測できる．さらに，住所に含まれないランドマーク名などを抽出するために，(α-3) 全く同じ住所を持つ複数の店舗名称に含まれる文字列を抽出する．たとえば同じ住所を持つ 2 つの店舗の名称に「六本木ヒルズ」という文字列が含まれていれば，ビルの名称であると考えられる．

以下，2 で既存手法について簡単に説明し，3 で店舗名称に含まれる不要語のパターンを示す．4 で開発した手法の詳細を述べ，5 で実験による処理精度の検証結果を示す．6 で関連研究と比較し，7 でまとめる．

## 2. 店舗情報検索手法

### 2.1. 店舗情報検索における店舗名称の役割

われわれが提案した店舗情報検索手法では，ある実世界上の店舗 $A$ に関する情報を Web ページから抽出する場合，まず (p-1) 店舗 $A$ に関連するページを絞り込むための識別語（電話番号など）をキーワードとしてサーチエンジンに与えて Web ページを収集し，次に (p-2) それぞれの Web ページを HTML のタグ構造によってブロックに分割し，店舗 $A$ を識別する語（これも電話番号など）を含むものを取り出すという処理を行う [10]．p-1 で用いる検索キーワードや p-2 の識別語には，一意性が高く表現に曖昧さが少ない電話番号を用いることで高い適合率が期待できるが，一方で，電話番号が記載されていない Web ページを収集することができない．われわれは，店舗情報検索にとって店舗利用者による評価・評判情報（いわゆる「口コミ情報」）が非常に重要であると考えており，そのような第三者によるページには電話番号が記載されていない可能性が高いという仮説を立て，店舗の電話番号に加え店舗名称と住所も検索語，識別語として利用している．

実際に店舗に関連する Web ページの文章中に電話番号が含まれるかどうかは業種や地域によっても差があるが（たとえば予約が必要な美容院に関するページでは，第三者によるページでも電話番号が記載されていることが非常に多い），東京都内のレストランを対象に収集した Web ページからサンプリング調査を行った予備実験の結果では，人間が目視により口コミページであると判断したページの 21.7% には電話番号が含まれていなかった [11]．これらのページは店舗の名称と住所を用いて初めて識別できるため，検索キーワード・識別語として電話番号に加え，店舗名称と住所を用いることには十分な理由がある．ただし電話番号は単独で識別語として利用できるが，店舗名称が一致するだけでは支店や偶然同じ名前の店舗である可能性が，住所が一致するだけでは同じビルに入居している別のテナントの可能性があるため，店舗名称と住所の両方が含まれている場合に限ってその Web ページが店舗 $A$ に関する情報を含むと判断する．

### 2.2. 店舗名称の表記揺らぎによる問題

さて，実際に大規模な店舗データベースを用いて上記 p-1，p-2 の処理を行う場合，店舗データベースに登録されている正式な店舗名称（以下，単に「店舗名称」）と Web ページで用いられる店舗名の表記（以下，「店舗名表記」）が完全には一致していないために，検索精度が低下するという問題が起こる．たとえば，店舗名称が「食事処 第一食堂」である場合 [1]，第三者が作成する Web ページでは単に「第一食堂」と表記されることが多い．この時，p-1 で店舗名称をキーワードとすると Web 上に存在するページの一部が検索されず，再現率が低下する．また，p-2 で店舗名称が完全に含まれることを判定基準として用いた場合，多くのページが条件を満たせないためやはり再現率が低下する．

p-2 で再現率の低下を防ぐため，店舗名称と店舗名表記が部分的に一致していれば正解とする方法も考えられる．その場合には上の「第一食堂」は正しく識別されるが，「富士そば六本木ヒルズ店」と「ビックラーメン六本木ヒルズ」のように，全く異なる 2 つの店舗が共通する文字列を含み，かつ同じ住所を持つというケースは多数存在するため，適合率が大きく低下する．

そこで，店舗データベースに登録されている店舗名称をそのまま用いるのではなく，第三者が作成するページには記載されない可能性の高い，余分な部分文字列（不要語）を除去することにより，ジオパース処理の精度を向上させる前処理を考える必要がある．こ

1) 本稿で例として示す店舗名称は断らない限り全て架空のものであり，同名の店舗が実在しても無関係である．

の処理を本稿では「店舗名称のクリーニング」と呼ぶ．店舗名称のクリーニングは，店舗名称に含まれる不要語を抽出する問題と考えることもできる．

## 3. 店舗名称に含まれる不要語のパターン

NTT レゾナントが収集した全国の飲食店データベース（74,461 件）を調査し，店舗名称に含まれる不要語を以下のように 5 つのカテゴリに分類した．

**(c-1) 会社種別**
「株式会社」「有限会社」など，会社の種別を示す文字列．

**(c-2) 業種種別**
「レストラン」「居酒屋」「焼肉」「中国料理」など，業種を示す文字列．

**(c-3) 一般的な本支店名**
全国的に広く用いられる本支店名を表す文字列で，「東口店」,「西口店」,「北口店」,「南口店」,「駅前店」,「本店」など．

**(c-4) 地名を用いた支店名**
「札幌店」「銀座店」「恵比寿店」「渋谷宮益坂店」など，地名を冠した支店名を表す文字列．

**(c-5) ランドマーク，ビル名を用いた支店名**
「なんばウォーク店」「アクアシティお台場店」など，ランドマーク名やビル名などの固有名詞を冠した支店名を表す文字列．

## 4. 提案手法

提案する店舗名称のクリーニング手法は，普遍的に用いられる不要語に対しては辞書を構築し，場所に特有な不要語に対しては辞書を構築できないため店舗データベースに含まれる情報を用いて対応する．

### 4.1, 不要語辞書の構築

不要語辞書には，3 で示した不要語の分類のうち，c-1 の会社種別と c-2 の業種種別，c-3 の一般的な本支店名に該当する文字列を手動で拾い出し登録する．これらの不要語は全国的に広く用いられるもので，業種の細分化や新語の登場などによって多少の増加はあるものの，件数も十分に辞書を維持できる範囲である．

実際に構築した不要語辞書に含まれる件数は Table 1 の通りである．

Table 1. 不要語辞書のサイズ

| 不要語の種類 | 件数 |
|---|---|
| c-1 会社種別 | 4 |
| c-2 業種種別 | 87 |
| c-3 一般的な本支店名 | 6 |

不要語辞書を用いた店舗名称のクリーニング手法のアルゴリズム α-1 は次の通りである．

**α-1. 不要語辞書を用いた不要語の抽出アルゴリズム**
1) 形態素解析手法と区切り文字 ( 空白など ) を用いて店舗名称文字列を単語列に分割する．
2) 各単語を不要語辞書から検索し，一致するものがあれば不要語とする．また，不要語に「店」「亭」などの本支店を表す語尾が付属した文字列も不要語とする．
3) 全ての単語に対して 2) の処理を繰り返す．

α-1 により，たとえば「びっくりカレー南口店」のような店舗名称から「南口店」が抽出できる．

### 4.2. 住所情報を利用した不要語の抽出

次に，3 で示した分類のうち，c-4 地名を用いた支店名に該当する不要語を抽出する手法について説明する．地名は数が多く（市町村のレベルでも三千以上），しかも時間とともに変化するため，辞書の維持が難しい．地名データベースから自動的に辞書を作ることも考えられるが，地名が店舗名の重要な一部として用いられることもあるため，不要語辞書に登録して単純に取り除くことはできない（たとえば「富士そば」という店舗名称に含まれる「富士」は地名だが，不要語ではない）．そこで，次のアルゴリズム α-2 を用いて不要語の抽出を行う．

**α-2. 住所文字列に含まれる不要語の抽出**
1) 対象店舗の住所を店舗データベースから検索する．
2) 形態素解析と区切り文字（空白など）を用いて店舗名称文字列を単語列に分割する．
3) 各単語が住所文字列に含まれるかどうか検査し，含まれる場合は不要語の候補とする．
4) 不要語の候補に「店」「亭」などの支店を表す語尾が続いた場合，この文字列（語尾を含む）を不要語として抽出する．
5) 全ての単語に対して 3)，4) の処理を繰り返す．

具体的な例として，店舗名称が「富士そば渋谷店」という場合を考える．上述したように，地名を全て抽出すると「渋谷」だけではなく「富士」も抽出されてしまう．しかしこの店舗の住所が「東京都渋谷区渋谷 × - △」であることを利用すれば，「渋谷」は支店名だが「富士」はこの店舗の場所から考えて支店名ではないことが推測できる．ただし，住所に含まれる地名を全て不要語とすると「北海道」のように店舗名に広く用いられている語まで抽出されてしまうため，支店

Table 2. A3 によって自動生成された住所とランドマーク・ビル名の組（一部）

| 住所 | 不要語集合 |
|---|---|
| 大阪府大阪市中央区千日前一丁目 | なんばウォーク店 |
| 東京都豊島区東池袋一丁目 29 番 | 池袋 60 階通り店 |
| 埼玉県川越市新富町二丁目 12 番 | 川越クレアモール店 |
| 東京都中央区京橋一丁目 1 番 | 東京駅前店 |
| 東京都港区台場一丁目 7 番 | お台場店 / アクアシティお台場店 |
| 神奈川県横浜市西区南幸一丁目 5 番 | ジョイナス店 / 横浜店 / 横浜ジョイナス店 |
| 福岡県福岡市中央区天神一丁目 4 番 | 博多大丸店 / 大丸店 |
| 大阪府大阪市北区梅田一丁目 12 番 | 梅田イーマ店 |

を表す語尾（「店」など）が続く場合だけを不要語とした．また，住所にビル名が記載されている場合には，3 で示した分類のうち **c-5** ランドマーク・ビル名を用いた支店名に該当する不要語も **α-2** で抽出できることがある．

### 4.3. 複数の店舗情報を利用した不要語の抽出

次に，3 で示した分類のうち **c-5** ランドマーク・ビル名を用いた支店名に当たる不要語を抽出することを考える．住所にランドマーク名やビル名が記載されていればアルゴリズム **α-2** で抽出できるが，実際には記載されていないことも多い．このようなランドマーク名は開発にともない年々増えるため，手動で辞書を維持することが特に難しい．そこで，ランドマークとなるような大きな施設には複数の店舗がテナントとして入居していることに注目し，同じ住所を持つ複数の店舗名称に共通する文字列をランドマーク・ビル名として抽出するアルゴリズム **α-3** を用いる．

**α-3．複数の店舗名称に含まれる不要語の抽出**

1) 対象店舗 $s_0$ と同じ住所（番・地番レベルまで）に存在する全ての店舗 $s_i$（$i = 1 .. n$，$n$ は同じ住所を持つ店舗数）を店舗データベースから検索する．
2) 各店舗 $s_i$ の店舗名称を形態素解析と区切り文字を用いて単語列 $w_{ij}$（$i = 1 .. n$，$j = 1 .. m_i$，$m_i$ は $s_i$ の店舗名称を分割した単語数）に分割する．
3) 1 つの店舗が複数の名前で重複登録されているケースがあるため，電話番号を用いてチェックする．店舗 $s_x$ と $s_y$ の電話番号が同じ場合，$w_x := w_x \cup w_y$ とし，$w_y$ を削除する．
4) 全ての $w_{ij}$ に現れる語をカウントし，2 回以上現れる語を不要語として抽出する．

たとえば「アクアシティお台場店」のようにランドマーク名が含まれる支店名では，アルゴリズム **α-2** を用いて不要語を抽出すると「台場店」だけが抽出され「アクアシティ」が残ってしまう．「六本木ヒルズ店」の場合には「六本木」が不要語候補となるが，その直後に「店」などの語尾が続かないため不要語として抽出されない．一方，アルゴリズム **α-3** を用いると，たとえば「富士そば六本木ヒルズ店」「ビックラーメン六本木ヒルズ」の二店舗が同じ住所であった場合，共通する文字列である「六本木ヒルズ」がランドマーク名として抽出でき，支店を表す語尾を含む「六本木ヒルズ店」までが不要語として除去され，それぞれ「富士そば」「ビックラーメン」のようにクリーニングされる．

ところで，アルゴリズム **α-3** で得られた住所と不要語の組みは，その住所に存在するランドマークやビルのデータベースであり，対象とする店舗の業種やデータセットによらず再利用が可能である．実際に作成された例の一部を **Table 2** に示す．この情報を蓄積することにより，ランドマーク名やビル名の抽出精度が向上していくことが期待できる．

## 5. 実験による検証

### 5.1. 実験の設定と結果

以下の 3 つの実験を行い，提案手法の性能を検証した．

**実験 1 クリーニング性能の検証**

目的：提案手法によって不要語が除去できることを検証し，その成功率を調べる．

対象データ：NTT レゾナントが収集した全国の飲食店データベース（74,461 件）．

方法：(1) 店舗名称をそのまま使う，(2) アルゴリズム **α-1** のみ，(3) アルゴリズム **α-2** のみ，(4) アルゴリズム **α-3** のみ，および (5)**α-1**，**α-2**，**α-3** を組み合わせた手法 (**α-123**) の 5 種類の手法を用いて店

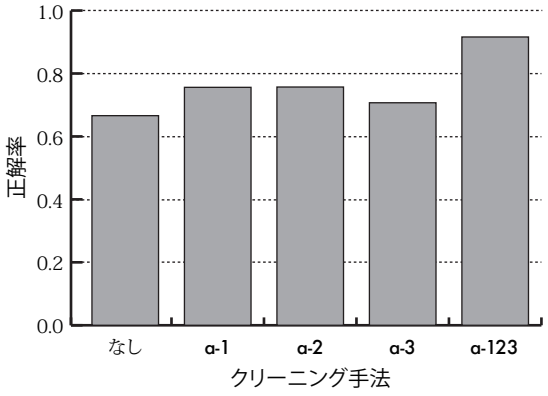


Fig.3 提案手法による正解率正解率の比較

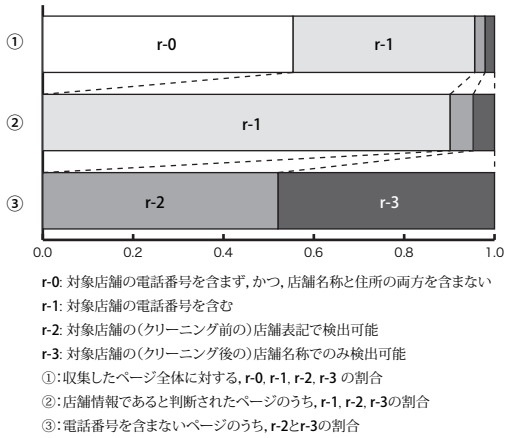


Fig.4 提案手法による Web ページ収集性能の改善

Table 3. 提案手法による正解率の比較

| 手法 | 正解件数 (988 件中 ) | 正解率 |
|---|---|---|
| なし | 658 | 0.666 |
| α-1 | 747 | 0.756 |
| α-2 | 748 | 0.757 |
| α-3 | 698 | 0.707 |
| α-123 | 905 | 0.916 |

Table 4. 収集した Web ページ数

| ページ分類 | 収集したページ数 | 割合 |
|---|---|---|
| r-0 | 394,074 | 0.5537 |
| r-1 | 286,261 | 0.4022 |
| r-2 | 16,309 | 0.0229 |
| r-3 | 15,023 | 0.0211 |

舗名称のクリーニングを行い, 正解率を比較する.

正解判定：対象データからランダムに 1,000 件サンプルを抽出し，電話番号が重複する 12 件を除いた 988 件について，手作業により支店名などの不要語を除去した正解セットを作成する．上記 5 種類の処理結果と，手作業によってクリーニングした結果が一致した場合に正解とする

結果:実験の結果を **Fig.3** および **Table 3** に示す．なお，対象データに対して **α-3** を適用した結果抽出された住所とランドマーク・ビル名は 1,768 組である.

**実験 2 店舗情報 Web ページ収集性能の検証**

目的：提案手法により店舗情報を含む Web ページの収集性能が向上することを示し, 増加率を調べる.

対象データ：実験 1 で用いた飲食店データベースを辞書として，[10] で示した手法により収集した，Web ページの集合

方法：対象データから電話番号を含まないものを選択し，クリーニング前の店舗表記でも収集可能なページ数と，クリーニング後の店舗名称でなければ収集できないページ数をカウントする

ページ分類：Web ページを以下のように分類する (1)**r-0**：電話番号を含まず，かつ，住所と店舗名称の両方を含まないページ数 (2)**r-1**：電話番号を含むページ数 (3)**r-2**：電話番号を含まず，住所と（クリーニング前の）住所表記を含むページ数 (4)**r-3**：電話番号を含まず，住所と（クリーニング後の）店舗名称を含むページ数

結果：実験の結果を **Fig.4** および **Table 4** に示す．増加率を $r_{improve}$=(**r-1** + **r-2** + **r-3**) / (**r-1** + **r-2**) と定義すると，$r_{improve}$ = 1.0497 である.

**実験 3 店舗ごとのページ収集性能の向上率の検証**

目的：提案手法により，Web ページの収集性能が大幅に向上する店舗が存在することを示す.

対象データ:実験 2 の結果, 提案手法により新たなページが得られた店舗（8,492 件）と，その店舗に関連づけられた Web ページの集合

方法：新たに得られたページが，店舗情報と判断されたページのうちに占める割合を店舗ごとにまとめ，その分布を調べる.

定義：店舗 $A$ に関連づけられたページのうち，電話番号を含むものの数を **r-1**($A$)，（クリーニング前の）店舗表記を含むものの数を **r-2**($A$)，（クリーニング後の）店舗名称を含むものの数を **r-3**($A$) とする．店舗によっては **r-1**($A$) + **r-2**($A$) = 0 となる（つまり，提案手法を用いない場合には 1 ページも収集できない）場合があるため，実験 2 と同様の増加率 $r_{improve}$ は利用できない．そこで，提案手法によって新たに得られたページが，その店舗

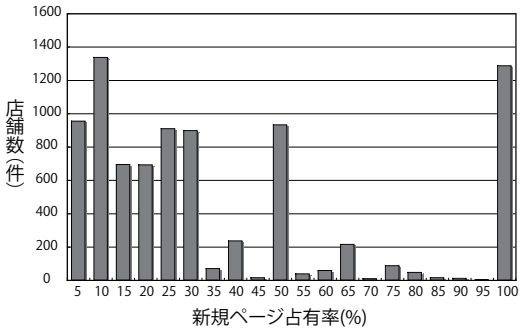


Fig.5 店舗によるページ収集性能向上のばらつき

の情報と判断されたページのうちに占める割合を新規ページ占有率と呼び，$r_{proportion}$ で表す．0 の場合は新しいページが全く得られなかったことを表し，0.5 の場合は提案手法により得られたページが半分を占めることを，1 の場合は提案手法を用いなければ 1 ページも収集できていなかったことを表す．

$$r_{proportion}(A) := \text{r-3}(A) \,/\, (\text{r-1}(A) + \text{r-2}(A) + \text{r-3}(A))$$

結果：実験の結果を Fig.5 に示す．

### 5.2. 考察

実験 1 の結果より，提案したアルゴリズム α-1，α-2，α-3 は，それぞれ 9%，9%，4% 程度の正解率向上に貢献していることが分かる．また，3 つのアルゴリズムを組み合わせた α-123 により，何もしなかった場合に比べ 25% 正解率が向上している．この値は 9+9+4=22(%) よりも高い値となっているが，これはたとえば「焼肉 炭火亭 六本木ヒルズ店」という店舗の場合にアルゴリズム α-1 によって「焼肉」が抽出でき，アルゴリズム α-3 によって「六本木ヒルズ店」が抽出できるというように，複数のアルゴリズムを適用することによって初めて正解となるケースが約 3% 存在しているためである．結果として，提案手法により店舗名称を 91.6% という実用的な高い精度でクリーニングできた．

一方，α-123 を適用しても正解とならないケースが約 8.4% 存在する．詳細な分析は今後の課題だが，ほとんどの場合，支店名に含まれている地名が住所に含まれていないことが原因で不要語が抽出できていない．地名と住所が一致しない原因としては，駅名が用いられている（「表参道店」, 住所は「渋谷区神宮前」），伝統的な地名が用いられている（「祇園店」, 住所は「博多区博多駅前」）の大きく 2 通りが見られた．

次に，提案手法の性能を「不要語の抽出精度」という面から検討する．何もしない状態で正解となる（すなわちクリーニングの必要がない）658 件を除いた 330 件に不要語が含まれているが，1 件に複数の不要語が含まれているケースもあるため（前述の「焼肉 炭火亭 六本木ヒルズ店」など）不要語の総数は 382 語であった．このうち，提案手法により 305 語を抽出することができたので，再現率は 305 / 382 = 0.798 である．また，不要語として抽出された 305 語を手作業で確認したところ，すべて不要語として適切であった（対象データに対しては適合率は 1.0）．

実験 2 の結果からは，提案手法によって有用なページを従来よりも 4.97% 多く収集できたと言える．約 5% という値は大きく感じられないが，これは全店舗での平均であり，店舗によってばらつきがある．実験 3 では，収集できるページ数が大幅に増える店舗が存在することを確認した．Fig.5 より，1,200 以上の店舗で新規ページ占有率が 100％，すなわち提案手法を導入することによってはじめて Web ページが収集できたことが分かる．ページ数が倍以上に増えた（新規ページ占有率が 50% 以上）の店舗まで含めると 2,700 以上あるが，店舗名称のクリーニングを行わない場合，これらの店舗では店舗表記に支店名などが含まれているというだけの理由で十分な数の Web ページが収集できない．さらに，ページ数に基づくランキングを行う場合にも順位が下がってしまうといった不公平が生じる．本実験の結果から，提案手法によりこれらの店舗に関連する Web ページの収集性能が向上し，問題が軽減できることが示された．

## 6. 関連研究

Web から実世界の施設に関するページを収集するという点で類似する関連研究として，病院などの施設名をキーワードに用いて Web から複数のページを収集し，その施設の住所を抽出する手法に関する研究が挙げられる（佐藤 [12]）．この研究では，情報統合の難しさとして同名の施設が存在する可能性や，施設の名称や住所の表記に含まれている揺らぎにより同一性判定が完全には行えないことを挙げている．さらに，これらの問題を吸収するための「属性の識別能力に基づく同一性判定」というアプローチを提案し，Web から収集した多数の類似した施設名・住所・電話番号の組から，複数の正しい施設名・住所・電話番号の組み合わせを求める部分を中心に議論している．

一方，本研究では，求めたい施設名・住所・電話番号の組み合わせは既知である（店舗データベースに存在する）という前提で, その施設に関連する Web ページを可能な限り多く収集することを目的とし，最適な検索キーワード・識別語となる名称を求める手法を議

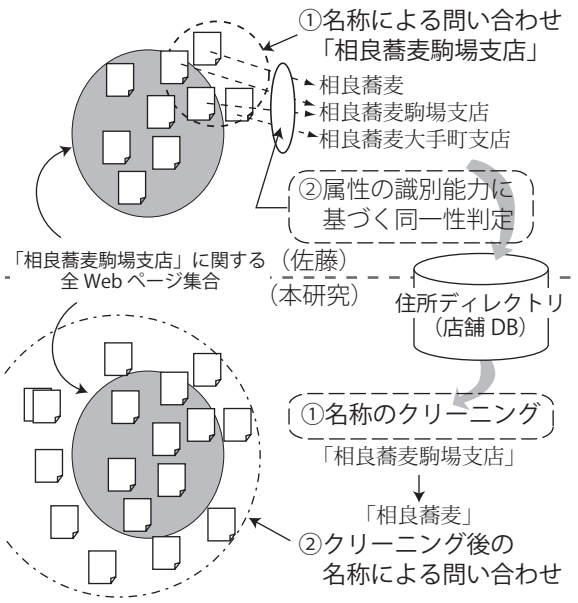


**Fig.6 関連研究との比較**

論している．

以上より，佐藤の研究は Web ページから店舗 DB を整備する手法であり，本研究は店舗 DB を用いてより広く Web ページを収集する手法であると捉えると，相補的な関係にあると言える (**Fig.5**)．また，佐藤の研究でも本研究と同様に「名称と住所の直積」に識別能力を認めているが，名称や住所の完全な同一性判定は不可能であるとして，クリーニング（表記の標準化）については触れていない．

## 7. おわりに

Web から店舗情報を収集する際に，検索キーワード・識別語として用いる店舗名称に含まれる不要語を除去するクリーニング手法を提案した．提案手法は，十分に小さな不要語辞書を用いる手法，店舗の住所を用いて支店名に含まれる地名部分を抽出する手法，および，同一住所に存在する複数店舗の名称に共通する文字列からランドマーク・ビル名を抽出する手法の 3 つを組み合わせた点が特徴であり，高いコストをかけて大規模な不要語辞書を整備する必要がないというメリットがある．

実験により，提案手法によって 90% 以上と高い精度で店舗名称をクリーニングできることを示した．また，店舗情報を含む Web ページの収集性能が全体で約 5%向上することも確認した．さらに，提案手法により新たに収集されるページの数は店舗によって大きなばらつきがあり，一部の店舗に対してはページ収集性能を大きく改善できることを示した．

今後の課題として，提案手法でも抽出できなかった不要語を分析し，(**i-1**) 手法を改良してさらに正解率を上げることと，(**i-2**) 提案手法が飲食店以外の業種でも有効であるか検証することが挙げられる．また，店舗名称のクリーニングによって，(**i-3**) 新たに得られた Web ページに特徴的な傾向があるか（たとえば口コミページが多く含まれているかどうか）についても検証する必要がある．

## 文　献

[1] Google Local, http://local.google.com/
[2] Yahoo Local, http://local.yahoo.com/
[3] AOL Local Search, http://localsearch.aol.com/
[4] MSN City Guide, http://local.msn.com/
[5] goo 地域, http://machi.goo.ne.jp/
[6] Einat Amitay, Nadav Har'El, Ron Sivan, Aya Soffer, "Web-a-Where: Geotagging Web Content", SIGIR2004, pp. 273-280, Jul, 2004
[7] T. Tezuka, R. Lee, H. Takakura, and Y. Kambayashi, "Acquisition of landmark Knowledge from Spatial Description", Proc. of International Conference on Internet Information Retrieval(IRC2002), Koyang, Korea, 2002.
[8] 相良 毅，有川 正俊，ジオパースによる Web からの空間コンテンツ獲得，電子情報通信学会 15 回データ工学ワークショップ (DEWS2004), I-11-01, Mar, 2004.
[9] 店舗情報検索エージェント実験グルメ版, http://labs.goo.ne.jp/agent/gourmet/
[10] Takeshi Sagara, Masaru Kitsuregawa, "Yellow Page driven Methods of Collecting and Scoring Spatial Web Documents", Workshop on Geographic Information Retrieval SIGIR 2004, pp. 4-8, Jul, 2004.
[11] 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, Nov, 2005.
[12] 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.