住所情報を用いた店舗名称のクリーニング手法

A method to normalize shop names using their address descriptions

相良 毅　　喜連川 優

Takeshi SAGARA, Masaru KITSUREGAWA

Shop search engine, which can retrieve shop information from the Web, is a growing topic in the area of Web search in this few years. Inside shop search engines, each crawled web pages must be checked whether it contains information of real-world-shops, such as restaurants or hotels, by shop name as an identifier. However, shop names are somewhat ambiguous; in some case, names in shop database are written in rigorous long form which will never used on web pages. This ambiguity makes difficult to extract shop names from web pages.

In this paper, we proposed a normalization method of shop names using their address, and other shop's name in the same building. The experimental results show that the method normalizes 91.6% shop names successfully, and increases the total number of web pages linked to shops by nearly 5%.

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[1-5]．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．この処理はジオパース(geo-parse)と呼ばれ，手動で行われていることも多いが，自動で行う場合には，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」に関連づけるか，地名や住所を抽出して実世界の「場所」に関連づける必要がある [6, 7]．

地域情報検索の中でも店舗情報は検索需要が高いが，対象となる Web ページ数が多いため，ジオパース処理を自動的に行う手法が不可欠である．そこでわれわれは，既存の電話帳を店舗データベースとして利用することで実世界に存在する店舗の情報を Web から収集し，場所に関連づける手法を開発した[8]．また，収集したページを店舗別に検索するシステムを開発し，店舗情報検索システムと呼んでいる[9, 10]．

店舗情報検索では，Web ページを各店舗に関連づける際に識別語の 1 つとして店舗名称を用いる．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）が含まれており，このようなノイズを含む名称を持つ一部店舗の場合に Web ページが十分に収集できない（再現率が低くなる）という問題がある．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（図 1）．

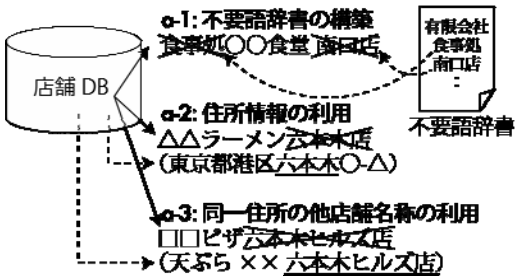

図 1. 提案手法の概要

Fig 1. Overview of Proposed Method

開発した手法の概要は以下の通りである．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，(a-1) 小さな不要語辞書を構築することで抽出できる．一方，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，(a-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，(a-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 で店舗名称に含まれる不要語のパターンを示す．3 で開発した手法の詳細を述べ，4 で実験による処理精度の検証結果を示す．5 で関連研究と比較し，6 でまとめる．

2. 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．

3. 提案手法

提案する店舗名称のクリーニング手法は，場所によらず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．

3.1, 不要語辞書の構築

不要語辞書には，2 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表1の通りである．

表 1.　不要語辞書のサイズ
Table 1. Size of Disused Words Dictionary

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 a-1 は次の通りである．

a-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．

a-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．

3.2. 住所情報を利用した不要語の抽出

次に，2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，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a-2を用いて不要語の抽出を行う．

a-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a-2で抽出できることがある．

3.3. 複数の店舗情報を利用した不要語の抽出

次に，2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．ランドマークやビル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a-3を用いる．

a-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「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 a-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表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．

4. 実験による検証

4.1. 実験の設定と結果

以下の 2 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

表 2. a-3 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)
Table 2. Extracted Pairs of Landmarks/Building Names and Their Addresses (part.)

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

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 a-1 のみ，(3)アルゴリズム a-2 のみ，(4)アルゴリズム a-3 のみ，および(5)a-1，a-2，a-3 を組み合わせた手法(a-123)の 5 種類の手法を用いて店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：実験の結果を表 3 に示す．なお，対象データに対して a-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.

表 3. 提案手法によるクリーニング成功率
Table 3. Precision Rate of Proposed Cleaning Methods

手法	正解件数(988 件中)	正解率
なし	658	0.666
A1 のみ	747	0.756
A2 のみ	748	0.757
A3 のみ	698	0.707
A1+2+3	905	0.916

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名称をクリーニングすることで，Web ページの収集性能の向上率を調べる.

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 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.

表 4. 収集した Web ページ数
Table 4. Number of Extracted Web Pages

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

4.2. 考察

実験 1 の結果より，提案したアルゴリズム a-1，a-2，a-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 a-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 a-1 によって「焼肉」が抽出でき，アルゴリズム a-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, a-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（対象データに対する適合率は1.0）.

さて，実際に店舗に関連するWebページを収集する際には電話番号や住所も店舗の識別子として利用する．名称が一致しただけでは同名他店の可能性があり，住所が一致しただけ

では同じビルに入っている他店の可能性があるため，住所と名称の両者が一致してはじめて対象店舗であると識別できる．そこで実験2では，これらの識別子を用いた場合の性能向上率を調べた．実験の結果から，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．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．店舗ごとに収集されたページ数の変化を調べたところ，提案手法を導入する前は１ページも収集することができなかったが，提案手法を導入することによってWebページが収集できた店舗が1,200以上存在した．このことから，店舗名称に不要語が含まれる店舗でWebページが十分に収集できないという問題が軽減されたといえる．

5. 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[11]）．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．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．

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．

[謝辞]
本研究は NTT レゾナント株式会社との共同研究「地理的情報を持つ評判情報のインターネットからの抽出技術の研究」の成果である．本研究の実施にあたり，店舗データベースの提供をはじめ，ご助力いただいたことに感謝する．

[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] Amitay, A., Har'El, N., Sivan, R., Soffer, A.,: “Web-a-Where: Geotagging Web Content”, SIGIR2004, pp. 273-280 (2004)
[7] Tezuka, T., Lee, R., Takakura, H. and Kambayashi Y.: “Acquisition of landmark Knowledge from Spatial Description”, Proc. of International Conference on Internet Information Retrieval (IRC2002) (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 (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] Sagara, T., Kitsuregawa M., “Yellow Page driven Methods of Collecting and Scoring Spatial Web Documents”, Workshop on Geographic Information Retrieval SIGIR 2004, pp. 4-8 (2004)
[11] 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)

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．1998 年東京大学工学部．1995 年東京大学大学院工学系研究科修了．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，地理学会会員．

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.