Geocrawler：個人サイトの評価情報と位置情報に基づいた店舗検索用Webインデクサの開発

新井 イスマイル† 川口 誠敬†
藤川 和利† 砂原 秀樹†

近年，口コミ情報サイトを例とする，ユーザの行動を基にした店舗・施設の検索サイトが注目されている．これらの検索サイトでは，位置に基づいた検索が可能であることと，店舗・施設に対して複数のユーザからの第3者の評価情報が取得できることが求められている．しかし，商用の検索サイトには広告収入や検閲の影響により，被評価店舗にとって不都合な情報が現れにくく第3者の評価情報の提供に問題がある．また，従来の情報取得手法ではWWW上の情報をすべて収集し，複雑な自然言語処理によって位置に基づいた評価情報を抽出する作業が必要となり，サービス構築コストが膨大となるという問題がある．そこで本研究では従来の全文型検索エンジンを活用し，目的の分野を示すキーワードと商用検索サイトを除外するキーワードを組み合わせることによって目的の第3者の評価情報を収集する手法と，単純な形態素解析と文字列のパターンマッチングを用いた文字列処理によって住所を抽出する手法を提案する．この手法に基づいてWebインデクサを評価した結果，一度の収集のうち44%が目的とする個人サイトであり，位置情報の取得再現率が59%という結果が得られた．

Geocrawler: Web Indexer for Store Search based on Geographical Information and Evaluation Information on Personal Web Sites

Ismail Arai,† Yoshihiro Kawaguchi,† Kazutoshi Fujikawa† and Hideki Sunahara†

A user expects that he/she can search stores and facilities from Web information space based on his/her behavior (Ex. Word-of-mouth communication sites). For this purpose, an appropriate information must be retrieved based on user's location. In addition, a user expect that he/she can retrieve actual impressions of other users against stores and facilities to decide his/her behavior. However, there are two major problems to achieve the above requirements. One is that the actual impression of other users are often omitted on the commercial web sites by the sponsor's claims. The other is that the cost for the information retrieval may become large because the existing search engines have to crawl most of Web sites and the complicated natural language processing have to be used. In this paper, we propose a new method which can obtain appropriate Web contents from Web search engines by inputting keywords that include user's objective information and black list information. In addition, the proposed method can extract the geographical information from the obtained Web contents by a morphological analysis and a simple pattern matching. As a result of evaluating the Web indexer based on the proposed method, 44% in all obtained Web contents conforms to user's objective. Also, the recall ratio of the extract of the geographical information is 59%.

1. はじめに

近年，インターネットを利用した情報検索の対象は多岐にわたり，これから向かう飲食店，専門店，旅行先などを決定するといった行動支援を目的としたインターネットの利用が注目されている．

このようなユーザの行動に基づいた情報検索では，ユーザの状況を考慮した情報の絞り込みと，提示された情報とユーザの嗜好の照合が重要となる．飲食店の検索を例にあげると，ユーザはぐるなび☆，タウンページ☆☆，じゃらん☆☆☆といった店舗や施設などの検索サイト（以下，商用検索サイト）を利用して，目的地や予算などを選択することによってユーザの状況に沿ったコンテンツに絞り込み，さらに写真や評価・評判な

† 奈良先端科学技術大学院大学情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology

☆ http://www.gnavi.co.jp/
☆☆ http://itp.ne.jp/
☆☆☆ http://www.jalan.net

どを参考にして自らの嗜好に即した店舗・施設であるか否かを判断する．商用検索サイトは情報の検索対象が明確に設定されていて，それに必要なデータ構造が設計されているため，目的地や予算といったユーザの状況に適した情報の提示に適している．

しかし，嗜好の適合を判断するための評価情報については情報提供者の性質に偏りがある．商用検索サイトの多くのコンテンツは雑誌の編集者などの専門家1人の意見が添えられることが多いため意見数が少ない．また，投稿形式を採用し複数のコメントを集めた場合でも，被評価者にとって不利益となる情報が記述されていた場合は商用検索サイトの運営者は被評価者から要請があった場合に情報を削除する場合も多く，第3者の評価者が評価対象に下したい評価（以下，第3者の評価情報）が提供されていない可能性がある．これらの第3者の評価情報をユーザに提供できないことは，コンテンツとユーザの嗜好との適合を判断する材料を不足させることになり，ユーザの嗜好に即した情報提供が実現できていないといえる．したがって，第3者の評価情報を検索者に提供できる機構が必要となる．実際にこのような第3者の評価情報は検閲の影響を受けない，個人が管理しその管理者のみが情報を更新する個人のHPやブログ（以下，個人サイト）に残るため，これらを情報収集対象とすることでユーザの要求を満たす情報検索システムが実現できる．

ただし，個人サイトでは情報の構造化は各個人に任されていて，満足な情報検索を実現するためにはインデックスの項目ごとに適切な解析手法を用意する必要がある．商用検索サイトを利用した情報検索の多くは行動する目的地を第1の絞り込み要因としていて，このような位置指向検索は行動支援の情報検索において欠かせない情報の絞り込み手段であることから，まず位置に基づいた評価情報のインデクシングを実現することが重要である．

現存のサービスを利用してユーザが嗜好に合った店舗情報にたどり着く手法として，まず商用検索サイトを利用して目的地周辺の店舗を絞り込み，その時の予算や気分に合った店舗を絞り込む．その後，各店舗に対する評判を確認するために全文検索エンジンに店舗名などをキーワード入力し個人サイト中の記述を吟味する方法が考えられる．商用検索サイトによって位置指向検索が実現され，手作業による個人サイトの検索によって第3者の評価情報を取得している．これらの作業を自動化できることはユーザにとって有益となる．

第3者の評価情報を適切に抽出するためには，Web上のコンテンツのうち商用検索サイトのものを除外し，なんらかの対象について評価・評判情報を記述していることを判定する必要がある．そのためには従来の技術では，Web上のすべてのコンテンツを収集したのちに，自然言語処理の係り受け解析を行うが，この手法はコンテンツ収集に膨大なストレージが必要となり，また特定ドメインに対して適切な係り受け解析を行うための機械学習作業が必要となるため，設備および時間に対するコストが非常に高い．一方，ユーザは全文検索によって店舗の評判を手作業で検索できたことを考慮すると，評価情報が記載されていると思われるWebページを提示するだけでもより効率の良い情報検索が実現できると考えられる．したがって，自然言語処理によって評判情報を抽出し，その情報を提供するサービスを構築する手法よりも簡易な手法によって評価情報が記載されたWebページを収集しその情報を提供するサービスを実現することが重要な課題となる．

また位置嗜好検索を実現するには，収集したコンテンツから位置情報を自動抽出する必要があり，従来の手法では形態素解析によって得られた品詞の組合せを実際の住所表記パターンと照合し，さらに比較用の住所辞書（全国の全住所記述を登録）と一致したもののみを抽出住所としたが，この比較用の住所辞書の構築コストは膨大である．したがって比較用の住所辞書の生成を必要としない住所抽出手法の実現が課題となる．

以上の議論をふまえて，本研究は，第3者の評価情報を提供でき，位置指向検索が可能な，店舗や施設などの検索サイトの構築を低コスト化することを目的とする．

2章では，第3者の評価情報を含むWebページの収集手法と，位置情報の抽出方法についての詳細な議論を行い，上述したようなシステム構築を低コスト化する手法を提案する．3章では，提案するシステムであるGeocrawlerの概要についてコンポーネントごとに機能と詳細を説明する．4章では，本システムの主要部分であるHTMLファイル抽出コンポーネントと住所抽出コンポーネントをラーメン店の店舗検索に適用し実験を行い，その性能を商用サイトの割合や再現率，適合率を求めることで評価し，今後の課題について述べる．最後に5章でまとめる．

2. 既存検索サイトの現状とサービス構築コスト削減手法の提案

店舗・施設検索において評価情報を取り扱った位置指向検索を実現するためには，適切な情報源の選択および情報収集方法について考慮する必要がある．本説

では「第3者の評価情報」を取得するために個人サイトを情報源として設定することがユーザの要求を満たすことを述べる．そして「位置指向検索」を実現するための手法について述べ，それらを低いサービス構築コストで実現する手法を提案する．

2.1 有効な評価情報の取得

利用者は個人サイト上で，掲示板・コメント書き込み・トラックバックを使い利用者間で幅広いコミュニケーションをとることができる．個人サイトは，匿名性があり，商用検索サイトで存在するような検閲がないことから，ある事柄，ある対象に対して何の遠慮もない意見を発信できる．個人サイトが発信する意見を収集対象にしている関連研究として，ブログを掲示板と同様の情報源と考え，評価表現抽出を利用した評判情報検索機能を持つシステムを開発している研究がある[1)~3)]．これらの研究において，個人が発する情報は社会的に影響力の強い口コミ情報との結び付きもあり，価値のある情報源としてとらえている．また Web ページ内の，ある対象に対しての評価情報の書き込みを，製品開発や企業活動に反映しようという試みもある[4)]．

そしてブログは，ホスティングサービスによる簡単な情報発信手段，時系列での情報確認，RSS（RDF Site Summary）[5)] による更新通知，トラックバックによる関連記事との相互リンクなどから登録者も増え続けている（2006年3月約868万人）[6)]．そのため，個人サイトのみを情報収集の対象としても，サービスを十分に機能する情報量を確保することが可能である．

既存の評価情報収集手法[7)] の実験中に，あらかじめ収集した blog を CaboCha[8)] によって係り受け解析を行い，評価表現の候補すなわち評価対象–属性–評価語の3つ組を抽出したものがある．例をあげると「HDD の容量が大きい」という文を解析した場合に「HDD–容量–大きい」といった3つ組が抽出される．精度の高い係り受け解析ができる一方，これらは SVM（Support Vector Machines）を利用することから機械学習のコストが問題となる．実験結果として抽出した対象–属性–評価の3つ組の200事例のうち，誤抽出があった事例が22.0%，形態素解析・構文解析誤りがあった事例が14.0%となっている．本研究では第3者の評価情報が記載された Web ページを収集することを1つの目的としているが，既存研究では blog のみが対象となっているため個人 HP による情報が損なわれる．また飲食店の情報に特化して評判情報を抽出しつつ上記の精度を維持するにはさらなるチューニングが必要となり，解析対象量が膨大であることから実サービスの実現が困難となる．このため，第3者の評価情報が記載された個人サイトの収集を簡易に実現できることが望まれる．

そのため，本研究では WWW 上にあるコンテンツの中から，商用サイトを特定しそれらを除外することによって個人サイトを見出し，特定の目的の情報検索において特徴的なキーワードを既存の全文検索エンジンにクエリとして送信することによって，第3者の評価情報が記述された個人サイトを収集する手法を提案する．google などの全文検索エンジンに検索語（ラーメンなどのグルメ情報に特化したキーワード），「住所」という検索語を入力することによって，飲食店に関する情報を実現する．検索語によってはレシピや食材に関する Web ページが抽出されてしまう可能性があるが，「住所」というキーワードが追加されることによって店舗の情報に絞り込む効果が発生する．また次節に述べる住所抽出対象の絞り込みの役割も同時に持つ．このキーワードの組合せに加えて，検索結果に大量に出現する商用サイトの名称を除外キーワードとして設定することによって個人サイトの出現割合を向上させる．以上述べたキーワードの組合せを全文検索エンジンに一度送信するのみの作業で第3者の評判情報を含む Web ページを収集することができれば既存手法の問題点であった自然言語処理による膨大な Web 文書の解析コストを削減することができ，効果的な解決手法となる．

一方，有効な評価情報が数多く集まる場には，公共の電子掲示板も考えられる．しかし，これらは書き込み量が増加するにつれ，1つの Web ページ上に現れる評価対象が増加し，収集した Web ページがどの店舗に対しての評価情報を示しているかを判断することが困難となるため対象外とする．

2.2 位置指向検索

ユーザの行動に基づいた情報検索では位置情報を取り扱うことが重要となる．徒歩で街中の飲食店を訪れるときに，駅から徒歩圏内の飲食店を探すといったように，ある基準点から行動可能な範囲の位置検索がユーザの直感に結び付きやすい．このような検索を実現するには検索対象に絶対位置情報を付加する必要があり，多くの地図アプリケーションではコンテンツに対して，緯度・経度を付加して地図上にコンテンツを表示している．したがって，コンテンツに対して手間なく緯度・経度情報を付加させることがサービス実現の鍵となる．利用者に現在地に基づいた位置指向検索を提供するには，まず Web ページから位置情報（住所，郵便番号，電話番号など）を抽出し，それをもと

に地図上に配置し提示する．

店舗情報が記載されたWebページには住所とともに固定電話の電話番号も記述されている可能性が高い．電話帳のデータベースがあれば，固定電話と住所の変換は可能となり住所抽出の補助が可能である．しかし，数字列をハイフンで区切る表記は電話番号以外にも多くあるため，正確なフィルタの設計は困難である．Webページを目視すると，住所が記述されている場所に対して，視覚的に近いところに電話番号が記述されている割合が高いが，HTML文書内での距離の同定は困難であるため，本研究ではまず住所文字列の抽出に注力する．

位置指向検索の関連研究として，Web上のHTMLファイルを収集し，そのファイルから位置情報（郵便番号，住所，駅など）を取得して緯度・経度に変換し，HTMLファイルを地理的な位置に配置して位置指向検索を実現しているものがある[9]．この研究では丁目レベルの全記述（例：東京都武蔵野市緑町3丁目）を約95万件，比較用辞書に登録して，これを形態素解析結果と照合することによって92%という再現率を実現したが，これを推移させて番地・号レベルの住所抽出をすることは辞書登録件数が膨大になり非現実的である．また区画整理による住所の変更についても，それに対する住所表記をすべて書き下ろして辞書の再構築をする必要があるためメンテナンスのコストが高い．

そのため，本研究では，選択的に収集した個人サイトのHTMLファイルを形態素解析にかけ，「地域」と分類された箇所を住所の可能性のあるものと判定し，その周辺に対してのみ住所のパターンマッチングを行うことによって，省作業な住所の抽出する．2.1節の提案手法によって得られるWebページを目視によって分析し，住所が記述されている箇所の「地域」形態素の連続数を把握し，その数に基づいたルール判定を行うことによってWebページ内の住所発見が可能となる．丁目・番地・号の記述は「地域」形態素と判定されないため，これらの形態素解析結果についても分析を行い，住所を完全に抽出するマッチングパターンを見出す．

そして抽出した住所を緯度・経度変換をし，Google Maps[10]上に表示させることで視覚的な位置指向検索を実現する．

3. Geocrawlerの概要

本システムは，個人サイトのHTMLファイルのみを収集し，個人サイト内に含まれる住所を抽出する．抽出した住所を緯度・経度に変換しGoogle Maps上に表示する．これにより視覚的な位置指向検索を実現する．**図1**に，本システムが実現する位置指向検索システム，Geocrawlerの処理の流れ図を示す．2章で述べたように本研究の提案事項は2つあり，さらに機能要素としては4つに分類される．今後の性能改善を機能要素ごとに独立して作業できることが望ましいため，機能要素に即して4つのコンポーネントからなるシステムを設計した．評価情報を含むWebページを選択的に収集するためのHTMLファイル収集コンポーネント，HTMLファイルを形態素解析し位置情報を抽出するための住所抽出コンポーネント，住所を緯度・経度に変換する緯度・経度変換コンポーネント，緯度・経度をもとにGoogle Maps上に評価情報を含むWebページのリンクを表示する情報表示コンポーネントといった4つのコンポーネントから構成される．以下に各コンポーネントの詳細を述べる．

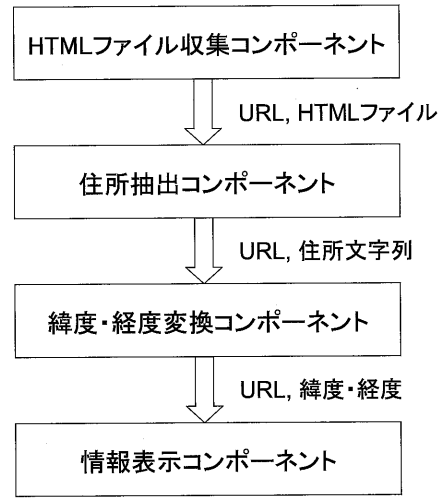


図1 処理の流れ図
Fig. 1 A flowchart of the proposed system.

3.1 HTMLファイル収集コンポーネント

HTMLファイル収集コンポーネントは，膨大なWebページから評価情報を含む個人サイトのHTMLファイルを収集する機能を持つ．

本研究では膨大なWebページを収集する作業を省略するために，Googleが提供するAPI（Google Web APIs[11]）を利用して，「ラーメン」といった料理名などのレストラン情報特有のキーワードを入力することによってWebページを収集する．

そして，商用サイトのURLを除外指定するキーワードを付加することによって，Googleから得られる検索結果を限りなく個人サイトのみの集合にする．具体的な商用サイトの定義については検索対象によって変

表 1 茶筅を使った解析例
Table 1 Analysis by Chasen.

形態素	分類
全国	名詞-一般
店	名詞-接尾-一般
)	名詞-サ変接続
八	名詞-数
ちゃん	名詞-接尾-人名
ラーメン	名詞-一般
住所	名詞-一般
福岡	名詞-固有名詞-地域-一般
市	名詞-接尾-地域
中央	名詞-固有名詞-地域-一般
区	名詞-接尾-地域
白金	名詞-固有名詞-地域-一般
1-1-27	名詞-サ変接続
うどん	名詞-一般
そば	名詞-一般
TEL	名詞-サ変接続
092-521-1834	名詞-サ変接続

化し実サービスごとに異なるため，4.1 節に述べる．

得られた URL のリストから，HTTP などを用いて Web 上のデータをダウンロードするツールである wget[12] を用いて HTML ファイルを収集する．

3.2 住所抽出コンポーネント

住所抽出コンポーネントは，HTML 収集コンポーネントによって得られた HTML ファイルから住所情報を抽出する機能を持つ．住所情報を抽出する機能を実現するために，形態素解析ツールの 1 つである茶筅[13] を利用する．解析を行う前に HTML ファイル内のタグは完全に除去する．HTML ファイル内のタグ構造（⟨td⟩ 奈良県奈良市三碓 2-1-10⟨/td⟩ など）を利用することによって，文書構造の解釈を補うことができる可能性はあるが，今回は個人サイトの HTML ファイルを住所抽出対象にしている．これら個人サイトの HTML ファイルは，商用検索サイトのものと違い，作成者のスキル・価値観によって構造が様々であり，タグの使用方法，使用箇所も様々となるため，タグ構造を利用しないこととした．

HTML からテキストに変換したファイルに対し，形態素解析した結果の例を**表 1** に示す．本研究では，形態素解析により「地域」に分類された形態素を住所の主要部分とする．まず予備実験を行った．HTML ファイル収集コンポーネントで収集した 178 ページの HTML ファイルをテキストファイルに変換し，形態素解析を行った結果を目視により確認した．その結果，「地域」形態素が 3 つ連続して出現した時点でその周辺の文字列は住所である可能性が高いことを確認した．「地域」形態素が 3 つ出現した後は「地域」形態素が続く限り形態素を抽出する．残りの丁目・番地

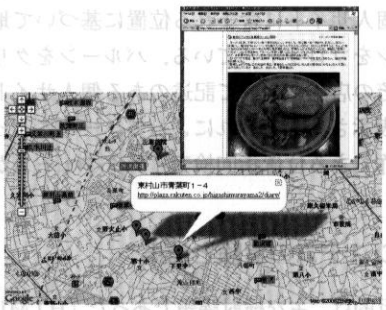


図 2 情報表示の様子
Fig. 2 Display image.

などを抽出するためにさらに余分に 5 つの形態素を抽出した後，住所記述のパターンを網羅する正規表現リストとのパターンマッチングを行い，住所文字列を抽出する．

3.3 サンプルアプリケーションの設計

上記の主提案に基づいて得られたインデックスを活用して Web アプリケーションを実現するために緯度・経度変換コンポーネントおよび情報提示コンポーネントを設計する．

緯度・経度変換コンポーネントは住所抽出コンポーネントで抽出した住所情報を緯度・経度情報に変換するジオコーダ機能を持つ．これについては Yahoo! Maps[14] を活用することによって実現する．Yahoo! Maps に住所を入力し検索を行うと，検索結果の URL 内に入力した住所に対応する緯度・経度が含まれている．検索結果ページの URL 部分の緯度・経度部分を自動抽出するスクリプトを作成することによって，緯度・経度変換コンポーネントを実現する．入力した住所が Yahoo! Maps の住所データベース内のデータと一致する場合は緯度・経度変換作業を行い，正しい緯度・経度を含む HTML ファイルを返す．しかし，入力した住所文字列が長すぎたり，短かすぎたりする場合は，Yahoo! Maps が自動的に近似（最長一致）の住所に対して緯度・経度変換される．したがって，緯度・経度変換コンポーネントの精度は，住所抽出コンポーネントの住所抽出の精度に依存している．

情報表示コンポーネントは，緯度・経度変換コンポーネントで得られた緯度・経度をもとに Google Maps 上にバルーン（マーカ）を表示する機能を持つ．緯度・経度変換コンポーネントで得られた数値（緯度・経度）を Google Maps API[15] に渡し，数値をもとに地図上にバルーンと個人サイトの URL を表示させる．これによって，個人サイトを対象とする位置指向検索が可能なグルメ検索サイトが実現する．実際にこのコンポーネントを利用した結果の表示画面を**図 2** に示

す．各個人サイトに含まれる位置に基づいて地図上にバルーンをプロットしている．バルーンをクリックすると，その店舗について記述のある個人サイトへのリンクが表示される．これによってユーザにとって直感的に操作できる位置指向検索インタフェースが実現できる．

4. システムの実験と結果

実験目的は，主な検討課題であった「HTML ファイル収集コンポーネントの個人サイト HTML ファイルを収集する精度」と「住所抽出コンポーネントの正しい住所を抽出しているかを評価する再現率と適合率」を確認することである．実験を行い，2 つのコンポーネントの精度を測定し，その結果について考察する．

4.1 HTML ファイル収集コンポーネントの実験と結果

HTML ファイル収集コンポーネントの目的は，全文検索エンジンのキーワード検索から得られる商用サイトと個人サイトの URL リストの中から，商用サイトの URL を除去し，個人サイトの URL をもとに HTML ファイルを収集することである．実験では，特定の検索語（Google に検索語「ラーメン 住所」を渡す）に加えて，指定ドメイン除外キーワードを付加する場合と付加しない場合の検索を行う．実際に Web ページでの住所記述にはほとんどの場合に「住所」といった項目名が添えられているため，検索語に含まれる「住所」というキーワードは，位置指向検索を可能にするために必要な要素として考える．そして，検索で得られた URL から HTML ファイルの内容を目視により確認し，個人サイト，商用グルメ検索サイトに分類する．

本システムでは，グルメ検索においての個人サイト（ブログ，個人 HP）と商用グルメ検索サイト（商用サイト）について以下のように定義した．なお，個人サイトの集合はブログと個人 HP の集合の非交和である．

- ブログ
  - RSS フィードを持つもの
  - トラックバック機能を持つもの
  - アーカイブによる過去ログ参照機能を持つもの
  - 時系列に日記が参照可能であるもの
- 個人 HP
  - ブログ，商用サイトの定義に該当しないもの
  - 店舗に対する評価情報を持つもの（★の数，数値の大小（4.5 点，8.0 点），言葉の強弱（普通，旨い，激旨）など，他店舗との違いを明確に表記していること）
- 商用サイト：
  - 店舗検索機能を持つもの
  - 全国規模の情報提供範囲を持つもの
  - 会員登録制機能を持つもの
  - サイト管理者が旅行会社，地域情報局，テレビ会社，新聞社であるもの
  - グルメ（ラーメン）以外の情報を提供しているもの（生活情報，コスメ，天気，交通など）
  - 上記のいずれかの内容を階層構造をたどることにより確認できるもの
- その他：
  - ネット通販，PDF，RSS 情報，リンク切れ

表 2 収集された Web サイトの内訳
Table 2 A breakdown of collected Web sites.

	商用サイト	個人サイト		その他	商用サイトの割合
		ブログ	個人 HP		
検索オプションなし	323	35	29	13	0.81
検索オプションつき	188	151	24	51	0.47

表 2 に上記の定義に従って分類を行った結果を示す．検索語：「ラーメン 住所」で検索を行った場合，検索結果上位 400 件中 64 件が個人サイト（ブログ 35 件，個人 HP29 件）であった．また商用サイトが 323 件であった．この結果は，商用サイトを除去し個人サイトの HTML ファイルを収集するという目的にそぐわない．個人サイトの情報を収集するために，検索結果の URL に出現頻度が高かった商用サイトに対して，以下のような除外キーワードを付けることで検索結果から除去する．

検索語：「ラーメン 住所 -ぐるナビ -Yahoo!グルメ -グルメウォーカー -all about -MSN グルメ -livedoor グルメ -ラーメンバンク -タウン -NAVITIME」の場合，上位検索結果 400 件中 175 件が個人サイト（ブログ 151 件，個人 HP24 件）であり，188 件が店舗情報，ラーメン総合案内サイトという集計結果が得られた．HTML ファイル収集コンポーネントを用いることで，HTML ファイル 400 件に対しての商用サイトの割合を 0.81 から 0.47 に減少させることに成功した．今後は除外キーワードリストの内容を充実させることによって商用サイトの除外率を向上させ，商用サイトの記述パターンなどを解析・学習することによって Web 文書の内容に基づいたフィルタリングを実現することが課題となる．

また，飲食店について第 3 者の評価情報が記述され

表 3 収集された Web サイトの記述内容
Table 3 Detail of collected web sites.

	飲食店情報		評価情報	
検索語	個人サイト	商用サイト	個人サイト	商用サイト
ラーメン	10	10	10	2
焼き鳥	10	10	6	0
うどん	10	10	8	1
パスタ	10	9	6	1
すき焼き	8	10	5	2
餃子	10	10	8	3
カレー	10	10	9	2
焼肉	10	10	6	1
ケーキ	10	10	7	0
平均値	9.8	9.9	7.2	1.3

ている Web サイトを収集する目的と照らし合わせて，収集される Web サイトに記述されている内容を**表 3** に示す．ラーメン店の取材本は数多く出版されていて，極端に有利なキーワードとなる可能性があるため，10 個の検索語を用いて検索し，個人サイト・商用サイトのそれぞれ上位 10 件の詳細を集計した．

飲食店について記述されている Web ページの 10 件中の平均数は個人サイトが 9.8 件，商用サイトが 9.9 件となり，おおむね収集できている．料理名を示す検索語に「住所」というキーワードを加えることによって飲食店に関する情報を的確に収集できたと考えられる．「すき焼き」というキーワードによって誤って収集された個人サイトの 2 件はすき焼き用の牛肉といった食材について説明されている Web ページであった．

評価情報が記述されている Web ページの収集数については個人サイトが平均 7.2 件となり，商用サイトについては収集数が平均 1.3 件となった．

収集された Web ページのうち個人サイトの割合は約 44%で，その中で評価情報が記載されている割合は平均 72%であったのに対して，既存手法の実験中での係り受け解析ではすでに収集済みの blog のみを解析対象として評判情報の抽出を行った結果，抽出した対象–属性–評価の 3 つ組の 200 事例のうち，誤抽出率が 22.0%すなわち正しい抽出率は 78%であった．前者は Web ページ単位で後者は文単位の統計量であることと，前者は客観的な評価情報を集計していることに対し，後者はすべての評価情報を集計しているため対等な比較はできないが，おおむね同等の割合で目的の Web ページが取得できることが確認できる．個人サイトの収集効率を省労力で実現することができれば，本手法は係り受け解析のための機械学習作業がまったく必要ないことからサービス構築コストが低く，ある程度の割合で目的の Web ページを収集することができる有効な手法であるといえる．

4.2 住所抽出コンポーネントの実験と結果

住所抽出コンポーネントの目的は，構造化されていない Web ページ上の文書から住所文字列を正確に抽出することである．実験は，まず HTML ファイル収集コンポーネントで収集した個人サイトの HTML ファイル 178 件を目視で住所確認し，正しい住所数を調査する．正しい住所の定義は，収集した個人サイトの HTML ファイル内に記述された住所が番地以降（「生駒市高山町 8916-5」，「三碓 1-5-10」など）まで記述された住所である．正しい住所の定義に従って，収集した個人サイトの HTML ファイルを目視した結果，769 個の正しい住所を確認できた．また住所抽出コンポーネントを用いて HTML ファイルから住所を抽出した結果，775 個の住所抽出に成功した．そのうち正しい住所は 457 個であった．正しい住所数と住所抽出コンポーネントを用いて HTML ファイルから抽出した住所数を用いて，住所抽出コンポーネントの再現率と適合率（精度）を以下に示す式を用いて求め，このコンポーネントの評価を行う．

$$\text{再現率} = \frac{A \bigcap B}{A} \tag{1}$$

$$\text{適合率} = \frac{A \bigcap B}{B} \tag{2}$$

A：目視により確認した正しい住所数
B：住所抽出コンポーネントで抽出した住所数

式 (1)，(2) を用いて，再現率と適合率を求めた結果，住所抽出コンポーネントの再現率は 59.0%，適合率は 59.0%になった．

住所抽出に失敗した主な要因には以下の 4 つがあげられる．

- 住所以外を抽出した例（95 件，49%）
  Web 文書中の住所の番地以降の記述と，電話番号の記述が連続するような状況などで，正規表現による文字パターンマッチングの照合箇所を間違えていることがあった．これは形態素解析によって解決できない問題のため，今後は数値文字列に対する意味づけを解決する手法について検討する必要がある．
- 大字，字，条が含まれた抽出例（40 件，20%）
  大字などの単語が「地域」形態素として分類されないことがあった．大字は住所で使われることが大半のため，茶筌の辞書に「地域」として新たに登録し，隣接単語の重み付けを調整することによって改善が図れる．

- 茶筅の辞書（20件，10%）
  「さいたま市」など，茶筅の辞書に存在しない地域名がいくらかあった．これについても，総務省の都市開発地域データベースなどを参照することによって，新住所キーワードを取得することができるため，茶筅に対して新たな辞書登録を行い，隣接単語の重み付けを調整することによって改善が期待できる．
- その他（41件，21%）
  主にHTMLファイルのタグ除去作業に失敗していることであった．改行コードを含まないWeb文書などでタグを除去することによって，想定しない形態素が生まれてしまった場合に間違いの原因となる．タグ除去後に，タグ内のキーワードの末尾に改行を付加することによって改善が期待できる．

以上の実験から既存手法に必要であった比較用の辞書を生成することなく，ある程度の再現率を確保した住所抽出手法が実現できたことを確認した．比較用の丁目まで記載された住所を登録した辞書は既存研究の実験当時で約95万件の事前登録が必要であった．本実験で明らかになったように最新の茶筌用の辞書を用いてもさいたま市などの地域が登録されていなかったことからWebサービスを構築・運用するにあたっては細かな辞書管理が必要となる．さいたま市には10の区があり，それに続く町字・丁目の組合せは膨大であることから既存手法のサービス構築コストは提案手法に比べて膨大であることが容易に想像できる．さらに番地・号の辞書登録をすることは困難であるため本提案は既存手法と比較してより低コストで実サービス化が可能な手法であるといえる．

今後は個々の要因に対してあげた改善策に取り組むことによって既存手法と同等の再現率を目指す．

5. おわりに

本研究では，第3者の評価情報を含むWebページを収集し，それらを位置指向検索可能とする情報検索サイトを低コストで構築することを目的として，第3者の評価情報を含むWebページを全文検索エンジンによって収集する手法と，単純な形態素解析と文字列のパターンマッチングを用いた文字列処理によって住所を抽出する手法を提案した．

第3者の評価情報を含むWebページの抽出をHTMLファイル収集コンポーネントによって実現し，このコンポーネントを使用する場合，使用しない場合の商用サイトを収集した割合を求め比較評価した．その結果，商用サイトが含まれている割合が81%から47%に減少した．また収集した個人サイトのうち他店舗との比較を明確に記述した評価情報は平均72%，商用サイトについては平均13%という結果を得た．これによって個人サイトを収集することが評価情報を収集することに結び付くことを確認し，既存手法において問題となっていた自然言語処理のための機械学習作業によるコストを削減したシステム構築手法を実現した．

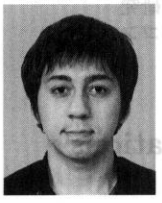
提案する住所抽出手法を住所抽出コンポーネントによって実現し，個人サイトのHTMLファイル178件から，このコンポーネントを用いて住所抽出を行い再現率と適合率を求め評価した．HTMLファイル178件から目視により得た住所数が769個，住所抽出コンポーネントを用いて抽出した住所数が775個，そのうち正しい住所数は457個であった．これらの住所数を用いて，再現率と適合率を求めた結果，再現率59.0%，適合率59.0%という結果になった．サービス構築コストが低い提案手法においても，ある程度の精度の住所抽出が可能であることを確認できた．抽出に失敗した内容を分析した結果，数値文字列に対する処理の強化および茶筌の住所辞書の強化を行うことで今後も性能に改善の余地があることを確認した．今後はこの課題を解決し，また適用範囲を飲食店全般さらには店舗全体へと拡大し，実用化に臨む．

参考文献

1) 鈴木泰裕，高村大也，奥村　学：Weblogを対象とした評価表現抽出，人工知能学会研究会資料 SIG-SW&-ONT-A401-02 (2004).
2) 新井イスマイル，飯田　龍，小林のぞみ，乾健太郎，藤川和利，砂原秀樹：グルメ情報を含むWeb文書からのユーザ指向型評判情報抽出システムの開発，情報処理学会，マルチメディア，分散，協調とモバイル（DICOMO2006）シンポジウム論文集，pp.953–956 (2006).
3) 武田英明：Weblog研究の現状，人工知能学会研究会資料 SIG-SWO-A402-06 (2004).
4) 松村真宏：チャンス発見のためのコミュニティマイニングに関する研究，博士論文，東京大学大学院工学系研究科電子工学専攻博士論文 (2003).
5) RDF Site Summary (RSS) 1.0. http://web.resource.org/rss/1.0/
6) ブログ及びSNSの登録者数（平成18年3月末）. http://www.soumu.go.jp/s-news/2006/060413_2.html
7) 鈴木泰裕，高村大也，奥村　学：Semi-Supervisedな学習手法による評価表現分類，言語処理学会第11回年次大会 (2005).
8) CaboCha. http://chasen.org/~taku/software/

cabocha/
9) 横路誠司，高橋克巳，三浦信幸，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.47, No.7, pp.1987–1998 (2000).
10) Google Maps. http://maps.google.co.jp/
11) Google Web APIs. http://www.google.com/apis/
12) GNU Wget. http://www.gnu.org/software/wget/
13) 茶筅．http://chasen.naist.jp/hiki/ChaSen/
14) Yahoo! Maps. http://www.google.com/apis/
15) Google Maps API. http://www.google.com/apis/maps/

(平成 18 年 10 月 31 日受付)
(平成 19 年 4 月 6 日採録)



新井イスマイル（学生会員）

平成 14 年明石高等工業専門学校専攻科機械・電子システム工学専攻卒業．平成 16 年奈良先端科学技術大学院大学情報科学研究科博士前期課程修了．現在，同大学情報科学研究科博士後期課程在学中．メタデータを活用した情報検索システムの研究開発に従事．電子情報通信学会，IEEE 各学生会員．

川口 誠敬

平成 17 年南山大学数理情報学部情報通信科卒業．平成 19 年奈良先端科学技術大学院大学情報科学研究科博士前期課程修了．現在，(株) NTT コミュニケーションズ．グルメ情報を中心に扱う Web アプリケーションの研究開発に従事．

藤川 和利（正会員）

昭和 63 年大阪大学基礎工学部情報工学科卒業．平成 3 年同大学院基礎工学研究科博士後期課程退学後，同年大阪大学基礎工学部助手等を経て，平成 14 年奈良先端科学技術大学院大学情報科学センター助教授，平成 17 年同大学情報科学研究科助教授，平成 19 年同大学情報科学研究科准教授，現在に至る．博士（工学）．分散処理システム，マルチメディアシステムの研究開発に従事．電子情報通信学会，IEEE，ACM 各会員．

砂原 秀樹（正会員）

昭和 58 年慶應義塾大学工学部電気工学科卒業．昭和 63 年同大学院博士課程修了．同年電気通信大学情報学部助手．平成 6 年奈良先端科学技術大学院大学情報科学センター助教授．平成 13 年同大学情報科学センター教授．平成 17 年同大学情報科学研究科教授，現在に至る．博士（工学）．インターネット，大規模広域分散環境，ネットワーク，並列処理，オペレーティングシステム，電子図書館に関する研究に従事．電子情報通信学会，ACM，IEEE 各会員．