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

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

奈良先端科学技術大学院大学

A Gourmet Search Engine Utilizing TPO Metadata and Reputation value Extraction Mechanism

Ismail Arai, Ryu Iida, Nozomi Kobayashi, Kentaro Inui,
Kazutoshi Fujikawa and Hideki Sunahara

Nara Institute of Science and Technology

1 はじめに

近年のネットワーク技術の進歩により，いつでもどこにいても情報にアクセスできる環境が実現されようとしているが，その一方でアクセス対象となるコンテンツは増加の一途をたどっている．ユビキタスネットワークの実現とともに，その膨大な情報にどのようにアクセスするかが重要な課題となる．近年ではWeb上のブログ記事に商品やサービスに関する有益な評判情報が記述されるようになり，その評判情報に対する企業や消費者のニーズが高まっている．

しかし，現実にはWeb上のテキストは膨大であり，検索された結果が真にユーザが必要とする情報かどうかを判定することが重要な課題となる．このような問題意識から，我々は有益な情報が記述されているブログ記事などからユーザ（検索者）の嗜好に合った評判情報を抽出する技術（ユーザ指向型評判情報抽出）の実現を目指す．

このユーザ指向型評判情報抽出の実現のためには，(1) ユーザが必要とする「時間」「場所」「状況」「検索目的」などに合わせて効率よく検索し，(2) 検索結果から評判情報に該当する箇所を適切に抽出する．という要素技術が必須である．

しかしながら，実用に耐えうるレベルの解析技術は未だ実現されていない．これら2つの技術を実用レベルで実現することによってユビキタス・コンピューティングを基盤とする高度情報化社会への貢献を目指す．本発表ではPDA (Personal Digital Assistants) を利用したグルメ検索シナリオを用意し，2技術連携の成果を報告する．

情報検索技術については，時間・場所・状況 (TPO) によって体系化された，ユーザおよびコンテンツの説明情報であるTPOメタデータを検索のクエリおよび検索対象とすることで，全文検索よりも効率的な検索を実現する．ユーザの状況に応じたグルメ情報を提供するために必要なメタデータの定義および評価式を提案する．またユーザのメタデータをあらかじめ設定しておくことによって省入力な検索も実現する．

評判情報の抽出については，文章内に記述された評判情報を〈検索対象，対象の属性，評価〉という形式で抽出する．これにより，ユーザが瞬時に評判情報を概観できる．またこの形式で抽出された情報をさらに要約することで，例えばレーダーチャートのように可読性の高い形で評判情報をユーザに提示することが可能になる．評価の候補となり得る表現を半自動で収集する手法を開発し，収集された表現を手がかりに，評価と評価の対象を同定する手法の開発に取り組む．特に研究対象をブログ記事に限定し，ブログ記事中の評判情報の抽出を試みる．

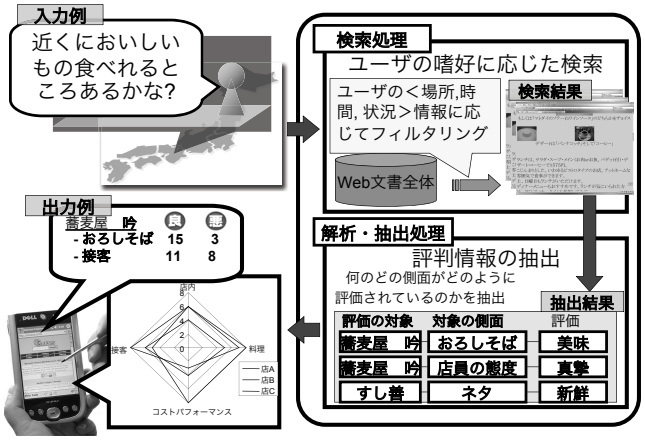

図 1: 評判情報抽出システム概要

第2章ではグルメ情報を含むWebブログ記事からのユーザ指向型評判情報抽出システムの開発目的と概要について述べる．第3章では検索エンジン部分および評判情報の抽出についての技術詳細を述べる．第4章では評判情報抽出システムの開発結果について述べ，第5章でまとめる．

2 評判情報抽出システム開発目的

WWW (World Wide Web) 上には膨大な量のコンテンツが蓄積されている．研究資料の検索やニュースの閲覧といった文書を対象とした調査が目的とされている割合が多いため，Google[5]

のような全文検索エンジンが広く使われている．近年，PDA や携帯電話の普及が高まるにつれて行動支援などを目的とした情報取得シーンが増加している．目的地への公共交通機関の乗換案内や，現在地周辺の天気予報の取得，飲食店の情報検索が挙げられる．全文検索では時間情報や位置情報をうまくキーワードに変換することができないため，目的のコンテンツにたどり着くことが困難である．現在はこのような行動支援目的のコンテンツを検索するには目的毎に用意されたポータルサイトを用意して，カテゴリツリーを辿る手法が主になっている．ただし PDA や携帯端末には時計や GPS といった現在の状況を取得するセンサを用意できるようになったことから，これらを活用してキーワードの選択を簡易化することが望まれる．W3C (World Wide Web Consortium) [8] にて議論されている Semantic Web[7] では情報検索のキーとして RDF (Resource Description Format) [6] 形式のメタデータを利用することによってインターネット上に存在する全ての Web データを機械処理可能とする構想がなされている．今後は検索者の状況を計算機環境が読み取り，検索クエリに反映できることが望まれる．

また，携帯移動端末は描画領域の制約があり，一度に確認できる情報量が限られている．検索のおおまかな流れは検索キーワードを送信し，検索結果を個々に確認して目的のコンテンツを得るという形式になる．検索結果を吟味する際にそのコンテンツの要所を見出すまでに全文を読む必要がしばしばあるが，これを携帯情報端末上で行うことは大きな労力となる．このような吟味の作業を計算機によって支援できることが望まれる．

これらの要求を満たすために，情報検索技術と自然言語処理技術を連携開発を行った．また行動支援の対象となるコンテンツと，意見が含まれているコンテンツの両性質が併存するコンテンツに対する効果が最も期待できるため，レストラン情報の検索シナリオに取り組んだ．

全体の概要は図 1 のようになる．ユーザは自らの状況を反映した検索クエリを検索エンジンに送信する．検索エンジンはクエリを解析し，ユーザの状況に適したコンテンツを整理する．そして上位の検索結果について評判情報を抽出して，検索者にはコンテンツ群の評判情報を要約したものを提示する．

3 各要素技術の概要

開発したシステムは検索技術と自然言語処理技術の 2 要素技術によって構成される．3.1 小節ではメタデータ検索技術の詳細について，3.2 小節では評判情報抽出技術について述べる．

3.1 M3 search engine

簡便に的確な情報検索を実現するにはメタデータを活用した情報検索が有効である．さらにはユーザの状況に対応したメタデータを活用することによってユーザの状況を反映した情報検索を実現できる．著者は以前より TPO (Time, Position, Occasion) によって体系化された TPO メタデータを活用した情報検索システムである M3 (Make the best use of Mutual Metadata) search engine[1, 2] を開発してきた．TPO メタデータはユーザの行動支援の対象となるコンテンツに対して有効となるため，本報告で特化したドメインであるレストラン店舗検索においても有効であると期待できる．

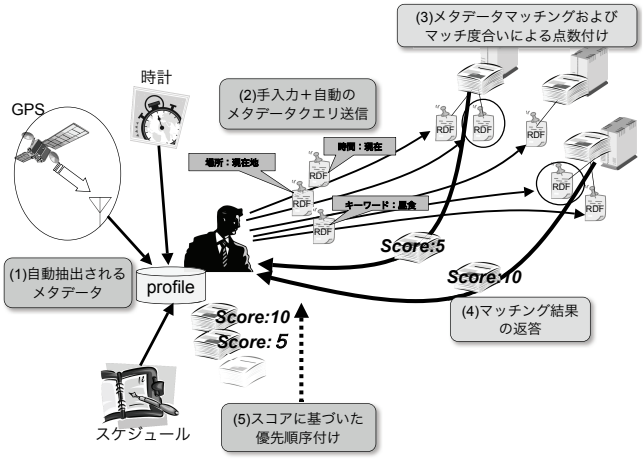

図 2: M3 サーチエンジンの動作概要

M3 search engine の動作概要を図 2 に示す．ユーザおよびコンテンツにはあらかじめ，それらの説明情報であるメタデータが付加されている．ユーザに付加されているメタデータをユーザメタデータ，コンテンツに付加されているメタデータをコンテンツメタデータと呼ぶ．ユーザメタデータは図 2(1) に示すように，センサーによって自動取得できるものや端末にあらかじめ登録されている属性情報を活用することによってユーザの検索キーワード入力の手間を削減する．

そのユーザメタデータと少量の手入力キーワードを組み合わせて，M3 search engine に対してクエリを送信する (図 2(2))．

図 2(3) はマッチングおよびスコアリング作業を示す．M3 search engine は送信されたクエリとコンテンツメタデータを照合することによって各コンテンツが検索者の状況に合っているか否かを判断する．これをメタデータマッチングと呼ぶ．合致したコンテンツはメタデータ毎に用意された計算式によってマッチ度合いであるスコアを算出すし，コンテンツの URL とスコアの対のリストを生成する．

そして生成された URL とスコアの対をまとめてユーザに返答する (図 2(3))．

最後にユーザは図 2(5) に示すとおり，スコアの高い順にソートされた検索結果一覧を取得して，各コンテンツを吟味する．

実用化に当たってはコンテンツのメタデータが潤沢に用意されていることが前提となるため，コンテンツメタデータのインデキシング技術についても取り組む必要があるが，本論文では 4 章で述べるとおり，簡便なインデキシングによってコンテンツメタデータを用意した．

3.2 評判情報抽出技術

M3 search engine によって膨大なコンテンツ中からユーザの状況にあったものにフィルタリングされるが，実際に情報検索をする上で，各コンテンツの吟味の作業は大きな労力となる．テキストから意見や評判などの情報を抽出・要約して提示することによって，より簡潔な情報検索が可能となる．

我々はこれまで，評価と評価されている属性の対で記述できる意見に着目し，機械学習ベースの抽出手法を開発してきた [4]．意見情報抽出の流れを図 3 にそって簡単に述べる．抽出手順は

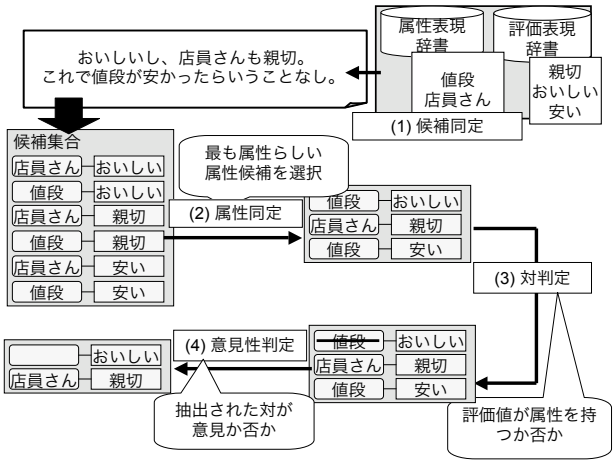

図 3: 意見抽出部分概要

(1) 候補同定，(2) 属性同定，(3) 対判定，(4) 意見性判定の 4 段階からなる．

評価，属性の候補の絞り込みは，あらかじめ作成した辞書を用いる (図 3(1))．次にそれぞれの評価の候補に対し，最も対になりそうな属性の候補を選択する (図 3(2))．ここで，図の「おいしい」の例にあるように，評価は必ずしも属性を持つわけではないため，評価の候補が選択された属性の候補と対になるか否かを判定する必要がある (図 3(3))．最後に，抽出された属性と評価の対が抽出したい意見か否かを判定する．何を意見とするかはアプリケーションにより異なるが，例えば「記述者の評価」を意見と考えると，図の「値段が安かったら」は仮定であって記述者の評価ではないため，意見とはみなさない．それぞれのモデル作成の詳細については，文献 [3] を参考されたい．

この 4 段階を通して，検索の結果得られた Web 文書から属性-評価の対を抽出する．抽出された意見が肯定的な意見か，否定的な意見かの判定については，今回はそれぞれの対もしくは評価表現に対し，肯定か否定か中立 (どちらでもない) かを人手で付与した辞書を作成して使用した．

4 評判情報抽出システムの実装

実際にレストランに関する情報を多く含む Weblog サイトを選択し，それらに含まれるコンテンツを加工して評判情報抽出システムを実装した．選択したサイトのコンテンツには投稿者が訪れたレストランに関する感想が Weblog 形式で蓄積されており，紹介された店舗の情報についても，店名，住所，電話番号，休日，営業時間が構造化されて記述されているため，本システムに必要となるコンテンツメタデータをスクリプト処理で容易に構築できた．

M3 search engine のスコアリングにおいて距離の計算が必要となるため，抽出した住所文字列を Yahoo! Map[9] に与えることによってリダイレクトされる URL に含まれる緯度・経度情報を抽出した．

インデックス化に成功したコンテンツは全部で 6891 件であった．そのうちのほとんどが東京都内のもので 4976 件あり，次に多かったのが大阪府内で 815 件であった．

```
<?xml version="1.0"?>
<contents>
    <RESTAURANT>
    <text>
    スープは醤油ベースで鶏がらからとったもので、あっさりとしつつ深みのある味です。
    麺は自家製の手打ち麺で超極細麺が特徴です。
    ツルツルとしたのどごしで、他店ではなかなか味わえない食感です。
    なんといってもおすすめは「ないすとラーメン」で、高山町伝統工芸の茶筅をかたち
    どったナルトが芸術的です。
    普通のラーメンを食べ飽きた方には特におすすめです。
    他にも自家製なら漬け、地鶏のチャーシューなど地域限定のメニューがふんだんに
    用意されています。
    近々石川方面に「じゃいすとラーメン」を出店する模様。
    </text>
    <FILEPATH>Nara/rvwid=17445</FILEPATH>
    <NAME>ないすとラーメン</NAME>
    <HOLIDAY>月曜（祝日の場合翌日）　</HOLIDAY>
        <HOUR>
            <li OPEN="11:00" CLOSE="14:00" />
            <li OPEN="18:00" CLOSE="23:00" />
        </HOUR>
        <ADDRESS>奈良県生駒市高山町8916−5</ADDRESS>
        <LATITUDE>34.43.40.220</LATITUDE>
        <LONGITUDE>135.44.07.904</LONGITUDE>
    <TEL>0743-72-0123</TEL>
        <ESTIMATE_DINNER>
            <LOW>1000</LOW>
            <HIGH>1999</HIGH>
        </ESTIMATE_DINNER>
        <ESTIMATE_LUNCH>
            <LOW>0</LOW>
            <HIGH>999</HIGH>
        </ESTIMATE_LUNCH>
        <GENRE>
            <li keyword="ラーメン" />
        </GENRE>
        <AIM>
            <li keyword="友人と" />
            <li keyword="家族で" />
            <li keyword="一人で" />
        </AIM>
    </RESTAURANT>
</contents>
```

図 4: コンテンツメタデータの例

検索に利用したメタデータは図 4 のようになる．一店舗は RESTAURANT 属性によって構成され，表 1 に示すとおりの 12 属性を含む．

HOLIDAY, HOUR, LATITUDE, LONGITUDE, ESTIMATE, GENRE, AIM 属性は M3 search engine のマッチングおよびスコアリング作業に用いられる．text 属性は評判情報抽出部分の解析対象となり，残りのメタデータは検索結果を表示するために活用される．

実際の動作画面を図 5 に示す．実際にデモを行う環境では時間や場所などが変化せず，時計や GPS 情報が有効に活用できないため，あらかじめプルダウンメニューでユーザメタデータの選択肢を用意した．左のトップ画面にてジャンル，時間帯，場所，予算，目的を選択して検索ボタンを押すことによって検索する．真ん中の画面は検索結果一覧を示している．最終的なスコアは肯定的な意見のカウントとした．これらの中から注目する店舗を選択することによって，右の画面のような店舗の詳細情報と意見の詳細を確認することができる．

このシステムを利用することによって簡便な操作でユーザの状況にあったコンテンツを検索することが可能となる．さらにコンテンツの平文を読むことなく，迅速に目的のコンテンツに到達することが可能であることが確認された．

本報告では肯定的な意見が上位になることが望ましい検索の一例をレストラン検索のターゲットに絞り込んで優位性を確認したが，逆に否定的な意見を抽出することが重要になるシーンもある．例えば企業が発売した商品に対する本音の評価が欲しいときに，否定的な意見というものはショッピングサイトから抽出することは難しい．迅速に特定の評価情報を収集するために本手法は有効であると考えられる．

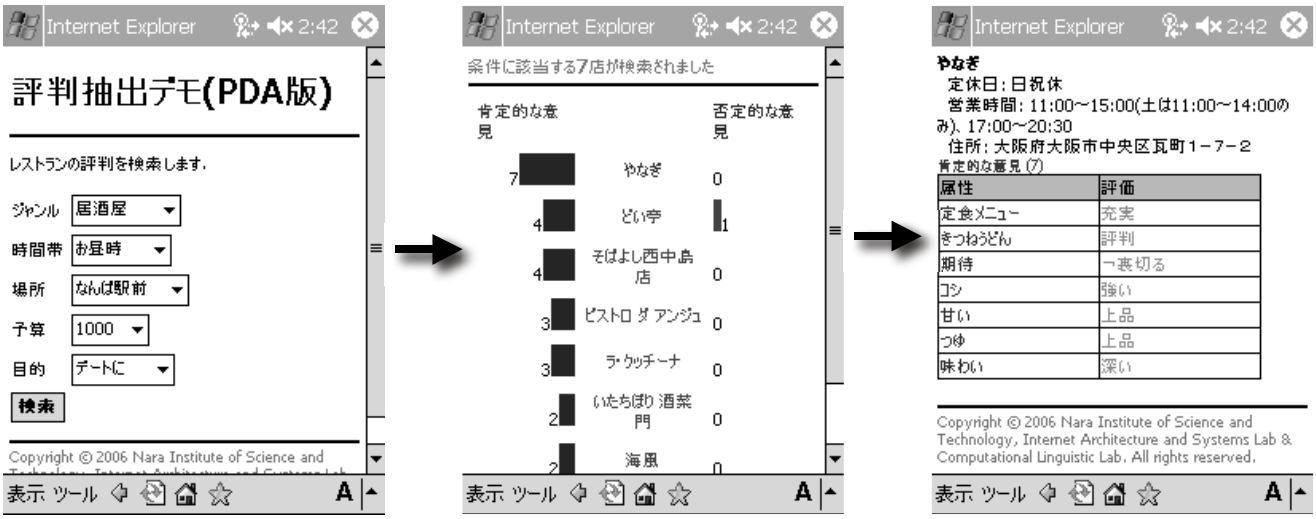

図 5: 動作画面

表 1: 設定したコンテンツメタデータ

属性名	説明
text	評判情報を抽出するための平文情報
FILEPATH	ローカルに保存したファイルパス
NAME	店舗名
HOLIDAY	休日
HOUR	開店時間
ADDRESS	住所
LATITUDE	緯度
LONGITUDE	経度
TEL	電話番号
ESTIMATE	予算
GENRE	ジャンル情報
AIM	目的情報

本報告ではインデックスの構築コストを下げるために，必要要素の揃ったサイトを限定したが実際の Web 文書は全てのメタデータ要素が構造化して用意されていることは希である．今後は WWW 上の無作為に発見した Web 文書に対してインデキシングを可能とするアルゴリズムを開発し，実用化に望みたい．

5 おわりに

WWW 上に膨大に存在する Web ページの検索を効率的に行うために，ユーザの状況に応じた検索と検索結果となる Web ページの要点抽出の 2 技術をかけあわせた評判情報抽出システムを開発した．ドメインをレストラン検索に絞り込み，稼働させた結果，プルダウンメニューから数個の選択を行うだけで行き先のレストランを決定できる情報検索システムを構築できたことを確認した．今後は対象ドメインの拡大に対応し得るインデクサの開発を行い，実用性を追求したい．

参考文献

[1] Ismail Arai, Kazutoshi Fujikawa, and Hideki Sunahara. A Proposal of information retrieval method based on TPO metadata. *Proceedings of 2005 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, Victoria, Canada*, pp. 442–445, August 2005.

[2] 新井イスマイル, 中村豊, 藤川和利, 砂原秀樹. TPO メタデータに基づく情報検索手法の提案. IPSJ, 第 12 回マルチメディア通信と分散処理ワークショップ, pp. 251–256, December 2004.

[3] 小林のぞみ, 飯田龍, 乾健太郎, 松本裕治. 照応解析手法を利用した属性ー評価値対および意見性情報の抽出. 言語処理学会第 11 回年次大会論文集, March 2005.

[4] 飯田龍, 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一. 意見抽出を目的とした機械学習による属性ー評価値対同定. 情報処理学会研究報告，自然言語処理研究会，2005-NL-165, pp. 21–28, May 2005.

[5] Google. http://www.google.com/.

[6] Resource Description Framework Model and Syntax. http://www.w3c.org/RDF/.

[7] SemanticWeb.org. http://www.semanticweb.org/.

[8] The World Wide Web Consortium (W3C). http://www.w3.org/.

[9] Yahoo!地図情報. http://map.yahoo.co.jp/.