ブログの評判情報を用いた施設情報検索

松本章代[†]　草桶慎太郎[†]　Martin J. Dürst[†]

Local Search Using Reputation on Blogs

Akiyo Matsumoto,[†] Shintaro Kusaoke[†] and Martin J. Dürst[†]

1. はじめに

現在，評判情報を含む，施設の情報を提供しているサイトの代表格としては，グルメ情報サイトの「ぐるなび☆1」や宿泊予約サイトの「楽天トラベル☆2」などが挙げられる．これらのサイトは情報量が豊富である反面，広告等の役割も兼ねており，基本的に施設・店舗側にとって有利な情報となっている．一方，投稿型グルメ情報サイトの「食べログ.com ☆3」など，口コミを主体としたサイトもあり，こちらはユーザ視点で書かれた評判情報を提供している．また，施設の位置を地図上で検索できる「Google マップ☆4」にも，レビューを投稿できる仕組みがある．しかし，これらのサイトで「投稿」を行うには会員登録が必要であるなど敷居が高く，投稿件数が少なくなりがちであるため，評価が偏ってしまう可能性がある．

そこで本研究では，様々な施設に関するユーザの率直な感想をブログから大量に自動収集し，それらを検索できるシステムを開発する．ブログから評判情報を正確に抽出し，客観的で鮮度の高い情報をユーザに提供することを目指す．利便性を考慮し，ブログから抽出した評判情報を Google マップと連携させ，GPS 付き携帯電話から利用できる施設情報検索サービスとして構築する．実際に街中で利用されることを想定している．

† 青山学院大学 理工学部
College of Science and Engineering, Aoyama Gakuin University

☆1 ぐるなび http://www.gnavi.co.jp/
☆2 楽天トラベル http://travel.rakuten.co.jp/
☆3 食べログ.com http://tabelog.com/
☆4 Google マップ http://maps.google.co.jp/

2. 関連研究

ウェブから評判情報を自動的に収集・分析する研究は，近年盛んに行われている．収集した評判情報を検索できるようにするためには，検索対象と評価表現を適切に結びつけてデータベースに格納する必要がある．立石ら[1] は，検索対象の語と評価表現の語が一定範囲内に含まれていた場合に，その対象物に対する評判情報であるとみなして抽出を行った．しかし，例えば「今日は，レストラン○○に行った帰りに，駅前のおいしいラーメン屋でも食事をした.」といった文の場合,「レストラン○○」の評価が「おいしい」と抽出されかねない．そこで奥村ら[2] は，検索対象と評価表現の間の係り受け関係を考慮する手法を提案した．係り受け関係を用いることにより，検索対象と評価表現の関係は正しく抽出できる．この他にも，小林ら[3] による共起表現を用いた手法や，藤村ら[4] による文節の n-gram を利用した手法が提案されているが，これらはいずれも 1 文中に検索対象の語と評価表現の語が含まれている場合に対し有効な手法である.「レストラン○○に行ってきました.」という見出しのブログの本文中に「△△を食べた．おいしかった.」と感想が書かれているようなケースは取りこぼしてしまうという問題点がある．

一方，森本ら[5] は，ウェブ上から施設と住所を自動抽出して施設検索システムを構築している．1 ページに複数の施設の情報が記述されている場合などにおいて，各施設について書かれている範囲を正しく特定することの難しさを指摘し，構造化された複数の情報が記載されているウェブページから情報を抽出する[6] 必要性について述べている．

そこで我々は，ブログの見出し構造に着目する．記

事の見出しとそのスコープ（記事本文の範囲）を特定することにより，これまでの類似研究より高い精度・再現率で評判情報を抽出することを目指す．

3. システム構成

本システムは，評判情報のデータベースを持つサーバと，携帯電話に搭載されるクライアント側のアプリケーションソフトで構成される．このアプリケーションソフトは，検索条件や位置情報を入力し，検索結果を地図上に表示するための入出力インタフェースの機能を持つ．

データベースの詳細については，4 節で述べる．

3.1 処理の流れ

施設情報検索サービスでは，ユーザが指定した「地域」にある，ユーザが重視したい「項目」の「評価値」を持つ「施設」を地図上に示す．

本システムの処理の流れを述べる．

(1) ユーザは携帯電話（クライアント端末）上で「地域」「検索対象」「項目」「評価値」の 4 項目（ただし「項目」「評価値」は省略可）を入力する．地域は GPS から取得することも可能である．

(2) 入力データはデータベースサーバに送られる．

(3) データベースサーバ上で評判情報の検索を行い，入力地域近辺の「施設名」「緯度・経度」「評判情報」をクライアント端末に返す．

(4) クライアント端末は，サーバから受け取った情報を，Google マップ上に反映させる．まず各施設の位置情報としてポインタのみを示し，そのポインタが選択されると，施設名と評価文などの詳細情報を表示する（図 1）．

「地域」とは「八王子」などの地名で，「検索対象」は「病院」といった施設の種類もしくは「ラーメン」といった施設に関連する言葉でも構わない．「項目」は「価格」「味」など重視したい項目（属性）を表す言葉，「評価値」は「おいしい」や「きれい」，「格安」といった評価を表す言葉である．

評判情報の具体的な検索手順については 4.4 節で述べる．

3.2 動 作 環 境

データベースサーバの OS は，Linux である．RDBMS には MySQL，日本語形態素解析・構文解析には CaboCha[7] を用いている．ブログの加工処理（記事見出しと本文の抽出など）は Ruby で実装している．

本システムは，クライアントアプリケーションの動作対象として，GPS を搭載した携帯電話を想定している．ただし今回，バランス Wii ボードと Google ストリートビューを連携させ，屋内でも模擬体験ができるようなデモ用システム（図 2）として実装する．

図 1 検索結果の出力例

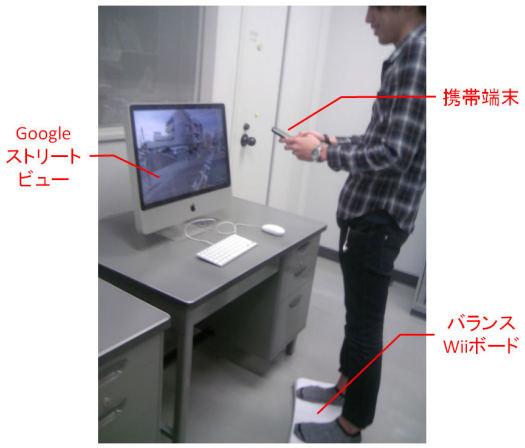

図 2 デモ用システム

4. データベースの構築

本システムは，施設テーブル，検索対象テーブル，評価文テーブルの 3 つから構成されるリレーショナルデータベースとして構築されている．全体のデータベース作成の流れを図 3 に示す．

4.1 評判情報データベースの作成

ウェブ上にある施設の住所リストとブログの評判情報とを結びつけ，評判情報に場所の情報を付加した形でデータベースに格納する．

データベースの作成に関わる各処理はすべて自動化されており，大規模なデータベースの構築が可能である．

表 1 施設テーブル

施設 ID	施設名	住所	緯度・経度
1	味の時計台	相模原市	35.540362, 139.431536
2	味の天徳	相模原市	35.558536, 139.374266
3	あじまる	相模原市	35.515270, 139.425879

4.2 データベースの作成手順

(1) Yahoo!電話帳☆1などから施設名＋住所の情報を抽出する．住所から緯度・経度を算出し☆2，施設テーブルを作成する☆3．

(2) 施設テーブルに登録した施設名と住所（市区町村名）を用いてブログ検索☆4を行う．上位 100 件ずつウェブ文書をダウンロードする．

(3) 記事見出しとその本文を抽出する．

(4) 記事の見出しに施設名が含まれるものを選ぶ．

(5) 抽出後の記事に対し，文章を抽出して各語の品詞の特定を行う．

(6) (5) の結果から名詞すべてを抽出し，検索対象テーブルを作成する．

(7) 検索対象テーブルと同様に，本文抽出後のウェブ文書から形容詞／形容動詞／連体詞／副詞を含む文すべてを抽出し評価文テーブルを作成する．

4.3 各テーブルの構成

各テーブルの属性を以下に示す．また，具体例を表 1・表 2・表 3 に示す．

- 施設テーブル
  - 施設 ID
  - 施設名
  - 住所
  - 緯度・経度
- 検索対象テーブル
  - 施設 ID
  - 記事 ID
  - 名詞リスト
- 評価文テーブル
  - 施設 ID
  - 記事 ID
  - 評価文

4.4 検 索 手 順

(1) 検索キーワードの「地域」で施設 ID を絞る．

(2) その中から，検索キーワードの「検索対象」または「項目」が含まれる記事 ID に絞る．

(3) さらにその中から，検索キーワードの「評価値」で記事 ID を絞る．

(4) その「評価値」が含まれる評価文に，その施設 ID から求めた施設名，緯度，経度を併せ，結果として返す．

☆1 Yahoo!電話帳 http://phonebook.yahoo.co.jp/
☆2 Google Maps APIs を利用．
☆3 現状では相模原市限定．
☆4 Google ブログ検索 http://blogsearch.google.co.jp/

表 2 検索対象テーブル

施設 ID	記事 ID	名詞リスト
1	1	相模原
1	1	北海道
1	1	ラーメン
1	1	国道
1	1	信号
1	1	味噌
1	1	豚骨
1	2	オープン
1	2	皆様
1	2	来店

表 3 評価文テーブル

施設 ID	記事 ID	評価文
1	8	ニンニクスライスかなり強く香ります
1	8	これが肉厚でなかなか旨いです
1	9	なかなか美味かったよ
1	10	麺は、ちょい太めのちぢれ麺
1	10	チャーシューは、脂身が多めだった

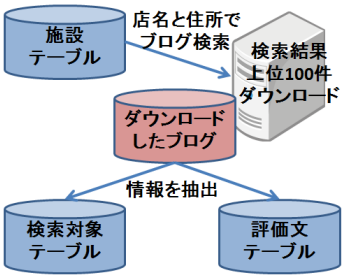

図 3 データベース作成の流れ

5. 適切な評価文を抽出するための工夫

適切な評価文の抽出するための手法について検討する．

5.1 記事見出しとスコープに基づく施設情報の抽出

ブログには，1 ページに複数の記事が存在することが少なくない．別の記事であれば，同じ施設について書かれている可能性は低いと考えられる．また，ページ内の非主要部分（広告など）に記載された情報も，記事の内容とは無関係である可能性が高い．つまり，これらを適切に判定し，自動切り分けを行わなければ，施設と評価文が正しく結び付かないことになる．

そこで，記事見出しに施設名が含まれているときに，そのスコープ内から評判情報を抽出する．たとえば 1 つのウェブページに図 4 の情報が含まれているとする．

```
<div>
  7 月 24 日（金）
  <h3>ラーメン○○へ行った．</h3>
  <div>
    ラーメンとギョウザを食べた．
    相変わらず量が多かった．
  </div>
</div>

<div>
  7 月 25 日（土）
  <h3>歯が痛い．</h3>
  <div>
    △△歯科医院に行った．
  </div>
</div>

<div>
  7 月 26 日（日）
  <h3>レストラン□□へ行った．</h3>
  <div>
    うなぎがうまかった．
  </div>
</div>
```

図 4 HTML サンプル

この場合，まず，3 つの記事が見出しと本文のセットになった状態で抽出される．3 つのうち，記事見出しに施設名が含まれているラーメン○○とレストラン□□の記事が選ばれる．ラーメン○○の記事からは，ラーメン，ギョウザ，量といった名詞が検索対象テーブルに格納され，相変わらず（副詞），多い（形容詞）を含む文「相変わらず量が多かった」が評価文テーブルに格納されることになる．レストラン□□の記事も同様に処理される．

このようなデータ抽出を実現するためには，ブログサイトによって異なるタグの使い方に対応しなくてはならない．すなわち，記事タイトルと本文を抽出するアルゴリズムは，ブログサイトごとに用意する必要がある．そこで，ブログ情報サイト☆で公表された，「2009 年 8 月のアクティブユーザー数☆☆が 10 万人以上のブログサイト」に該当する 10 サイト（アメーバブログ・FC2 ブログ・Yahoo!ブログ・livedoor Blog・JUGEM・ヤプログ !・goo ブログ・楽天ブログ・はてなダイアリー・seesaa）を現在日本における代表的なブログサイトとして解析対象に選んだ．

5.2 スパムブログの排除

ある調査によると，国内ブログの 4 割がスパムブログであるという[8]．我々はこれまでスパムメールの判別に関する研究を行っており[9]，これを応用した対策を検討中である．

☆ ブログファン http://www.blogfan.org/

☆☆ ブログファンによると，「毎月のアクティブユーザー数」とは，1 か月間のうちに更新されたブログのアカウント数．

6. む す び

ブログに掲載されている施設の評判情報からデータベースを作成し，GPS 搭載携帯電話から利用できるシステムを考案し，それをデモ用に実装した．

今後，抽出された評判情報の妥当性に関する評価実験を行う．実験結果から問題点を検証し，手法の改善案を検討する．

さらに，本システムの有用性を検証するため，携帯端末を用いて散策している状況を想定した被験者実験を行う．実験を通し，ユーザに対して検索意図に適った評判情報が提供できたかどうかについて，検証を行う予定である．

謝辞 本研究は文部科学省科学研究費補助金（若手 B，課題番号 21700116）の交付を受けている．

参 考 文 献

1) 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, 情処研報，2001-NL-144, pp.75–82 (2001).
2) 奥村学, 南野朋之, 藤木稔明, 鈴木泰裕：blog ページの自動収集と監視に基づくテキストマイニング，人工知能学会研究会資料, SIG-SWO-A401-01 (2004).
3) 小林のぞみ，乾健太郎，松本裕治，立石健二，福島俊一：テキストマイニングによる評価表現の収集, 情処研報, 2003-NL-154, pp.77–84 (2003).
4) 藤村滋，豊田正史，喜連川優：文の構造を考慮した評判抽出手法, DEWS2005, 6C-i8 (2005).
5) 森本泰貴，藤本典幸，長屋務，出原博，萩原兼一：Web を対象としたロボット型住所関連情報検索システムの開発, 信学論（D）, Vol.J90-D, No.2, pp.245–256 (2007).
6) Yanhong Zhai，Bing Liu：Web Data Extraction Based on Partial Tree Alignment, *Proc. 14th Int'l Conf. World Wide Web*, pp.76–85 (2005).
7) 工藤拓, 松本裕治: チャンキングの段階適用による日本語係り受け解析, 情処学論, vol.43, No.6, pp.1834–1842 (2002).
8) ニフティ株式会社: ニフティ、スパムブログのフィルタリング技術を開発, http://www.nifty.co.jp/cs/07shimo/detail/080326003337/1.htm (2008).
9) 藤田拓也，松本 章代，Martin J. Dürst: ベイジアンフィルタにおける言語知識を用いないトークン抽出方式の提案と評価，情処学論, Vol.50，No.9, pp.2182-2192 (2009).