様式Ｃ－１９

# 科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年　６月２５日現在

機関番号：３１３０２
研究種目：若手研究（B）
研究期間：2009～2011
課題番号：21700116
研究課題名（和文）ブログの評判情報を用いた施設情報検索およびストリートビューシステム

研究課題名（英文）Facilities Search and Google Street View Using Reputation in Blogs

研究代表者
松本　章代（MATSUMOTO AKIYO）
東北学院大学・教養学部・講師
研究者番号：40413752

研究成果の概要（和文）：本研究の目的は，施設評判検索システムの構築である．大量のブログから正確に評判情報を抽出するための手法を様々な角度から検討し，データベースを作成した．一方，クライアント側アプリケーションとして，GPS 機能付きの Android 端末（スマートフォン・タブレット）用の施設情報検索アプリを開発した．Google Maps 上に検索結果を（各評判情報とともに）表示することが可能である．

研究成果の概要（英文）：The purpose of this research is to build a reputation search system of facilities. We considered methods for accurately extracting reputation from a huge number of blogs and created database. Apart from the database, we also developed a client-side application. It is a facilities search application for Android smartphones and tablet devices equipped with GPS functions. It can show search results (with each reputations) on the Google Maps.

交付決定額

（金額単位：円）

| | 直接経費 | 間接経費 | 合　計 |
|---|---|---|---|
| ２００９年度 | 500,000 | 150,000 | 650,000 |
| ２０１０年度 | 400,000 | 120,000 | 520,000 |
| ２０１１年度 | 500,000 | 150,000 | 650,000 |
| 年度 | | | |
| 年度 | | | |
| 総　計 | 1,400,000 | 420,000 | 1,820,000 |

研究分野：総合領域
科研費の分科・細目：情報学・メディア情報学・データベース
キーワード：Web 検索・ブログ・評判情報・情報抽出・自然言語処理

## １．研究開始当初の背景

(1) 既存の施設情報サイトの問題点

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」や宿泊予約サイトの「楽天トラベル」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」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」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ため，投稿の敷居が高く，かつ情報が分散しているのが現状である．１つのサイトのみでは投稿件数が少なくなりがちであるため，評価が偏ってしまったり，古い

情報しかない，といったケースが少なからず見受けられる．

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 付きスマートフォン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．

(2) 評判情報元としてブログを用いる理由

評判情報の抽出にブログを用いる理由を 3 点挙げる．

まず 1 つ目は，ユーザの率直な意見が書き込まれている点である．ブログは様々な話題に対し個人の意見を自由に記述することができ，世間の意見や感想などの情報を知ることができる．

2 つ目は，ブログ利用者の増加である．総務省の調査資料によると，ブログ登録者は 2005 年 3 月末約 335 万人，2006 年 3 月末には約 863 万人と一般ユーザの利用が増加している[1][2]．また，総務省情報通信政策研究所の調査資料によると 2008 年 1 月、インターネット上で公開されている国内のブログの総数は約 1,690 万（記事数は約 13 億 5,000 万件）[3]．そして，毎月新たに開設されるブログ数は、近年、毎月 40 万から 50 万程度（新たに書き込みされる記事数は 4,000 万から 5,000 万件程度）で推移し、活発な情報発信が続いている．文章打ち込むだけで簡単に新しい記事を更新することができるため，新しい話題や新しい施設に対する情報，意見がすぐに記事として書き込まれる．そのため，多くの記事が存在することにより，より多くの意見を抽出できると考えられる．

3 つ目は，記事の中の施設情報（特にはグルメ情報）の多さである．NEC が運営する wisdom の記事によると，NEC 総研が一般ユーザはブログに何を書くかを調査したところ，グルメ情報（行ったお店などについて）が 25.8%（複数回答）であった[4]．このことからも，行ったグルメ施設やレジャー施設などの施設情報が多く抽出できると考えられる．

(3) 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[5]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[6]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[7]による共起表現を用いた手法や，藤村ら[8]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[9]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[10]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．

## 2．研究の目的

本研究の目的は，施設評判検索システムの構築である．大量のブログから正確に評判情報を抽出するための手法を様々な角度から検討し，データベースを作成する．一方，クライアント側アプリケーションとして，GPS 機能付きの Android 端末（スマートフォン・タブレット）用の施設情報検索アプリを開発する．Google Maps API を用いて，地図上に検索結果を（各評判情報とともに）表示する．

## 3．研究の方法

(1) ブログの実態調査の実施

適切な評判情報の抽出の実現するため，大規模なブログの実態調査を行う．具体的には，ブログの本文を構成する文・表・見出し・図という各構造に対し，評判情報がどのように記述されているかについて解析するシステムを構築し，実際に分析を行う．

(2) ブログの構造解析プログラムの開発

ブログから評判情報を適切に抽出するためには，ブログ内の記事見出しとその本文

を特定できることが必須である．そこでまず，ブログの構造解析を行うプログラムを開発する．これにより，見出しとそのスコープを特定することを可能にする．さらに，ウェブページ内の非主要部分を特定する手法を考案・実装する．

(3) 検索システムの構築

本システムは，評判情報のデータベースを持つサーバと，スマートフォン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．

これを以下の手順で構築する．

① データベース全体を設計し，各データ項目についてそれぞれ作成方法を検討する．
② 本システムの中心である，ブログから評判情報を適切に抽出する手法については特に詳細に検討し，実装する．
③ データベース全体を作成する．
④ クライアントの Android アプリを実装する．

(4) 評価実験

第 3 者の被験者に検索ワードを設定してもらい，検索結果（抽出された評判情報）の妥当性について評価実験を行う．実験結果から問題点を検証し，手法の改善案を検討する．

(5) スパムブログの排除

スパムブログの存在は，検索精度を低下させる要因となるため，これを排除する手法を検討し，実装する．

(6) 徒歩用ナビゲート機能の追加

システムの有用性を向上させるため，これまでに開発した Android アプリに徒歩用ナビゲート機能を追加する．これは，曲がるべき地点や進むべき方向を，音声あるいはバイブレータでアナウンスすることにより目的の施設までナビゲートする機能，また進行方向をカメラモード上に矢印で示す AR 機能の 2 つから構成する．

４．研究成果

(1) 検索システムの構築

システム全体の概要を図 1 に示す．

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．この評判情報データベースは，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．

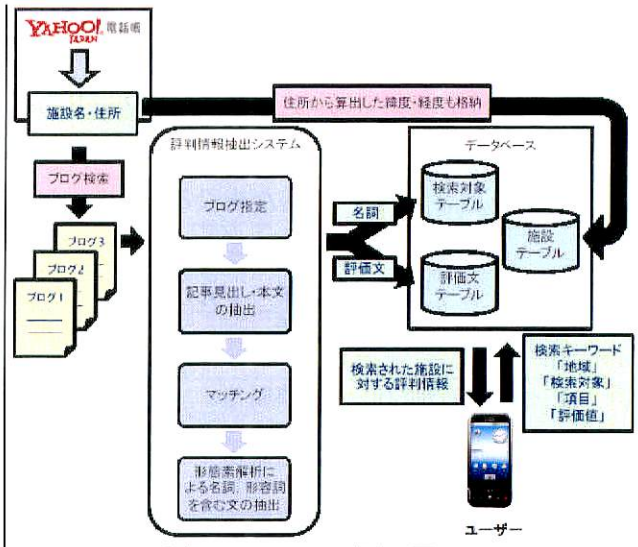


**図 1　システム概要図**

作成手順を以下に示す．

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

各テーブルの構成（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

検索時の処理の流れは以下のとおりである。

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．

**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 |

**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 |

② クライアントアプリの実装

実装した Android アプリの入出力画面を図 2 に示す．

このアプリは，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」を地図上に示す機能を持つ．「地域」とは「八王子」などの地名または GPS から取得した緯度・経度である．「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．


**図 2 入出力画面**

(2) 評価実験

① 実験方法

第三者の被験者に検索キーワードを作成してもらい，実際に検索したい評価文を取得することができたか，どのような問題があるかについて確認する．今回の実験では地域を「相模原」に限定し，検索対象を Yahoo!電話帳にある「ラーメン店」などの数種類の飲食店に対し検証を行う．検索キーワードの一部を表 4 に示す．

本実験の評価基準として，評価文をユーザが確認した時に明らかに検索内容と異なるものに対して不適合とし，それ以外を正しく検索されたものと判断する．そして，検索された全体の評価文の数と正しいと判断された評価文の数から検証する．

**表 4 検索キーワード**

| No. | 地域 | 検索対象 | 項目 | 評価値 |
|---|---|---|---|---|
| 1 | 相模原市 | ラーメン | スープ | 濃い |
| 2 | 相模原市 | ラーメン | スープ | まずい |
| 3 | 相模原市 | ラーメン | 値段 | 安い |
| 4 | 相模原市 | ラーメン | 野菜 | 多い |
| 5 | 相模原市 | ラーメン | チャーシュー | うまい |
| 6 | 相模原市 | ラーメン | 餃子 | おいしい |
| 7 | 相模原市 | ラーメン | 麺 | 太い |
| 8 | 相模原市 | ピザ | 宅配 | 早い |
| 9 | 相模原市 | ピザ | チーズ | おいしい |
| 10 | 相模原市 | 焼肉 | タン | 厚い |

② 実験結果と考察

表 4 に示された検索キーワードに対し，検索された全体の評価文の数と適切と判断された評価文の数を確認する．今回の検索キーワードの結果では全体の再現率は低くなり，精度は約 61% となった．

実験結果から 2 つの問題が挙げられる．1 つ目として，検索キーワードによって評価文

そのものの数が少ない場合や，存在しない場合が多くあるという再現率に関わる問題がある．この問題に対し，検索キーワードで指定する形容詞（評価値）の表現などによって検索結果が異なることが挙げられる．そして，様々な表現やひらがなや漢字などを変えて検索することで多くの評価文を確認することができた．また，別の原因として検索キーワードで指定する形容詞の活用によって検索結果が異なることが挙げられる．つまり，評価値に対して活用語尾を削除することで評価文の検索ヒット数が上がると判断できた．

2 つ目として，正しい評価文が多く検索できたもの，不適合であると考えられる評価文が多く検索できたものに分かれているという精度に関わる問題が挙げられる．これは，指定した項目に対して指定した評価値が使われていることが少なく，ブログの記事本文にその項目が含まれるが，ほとんど違う項目に対してその評価値は使われていることが原因であることが確認できた．これを解決するためには，評価文の判断を奥村ら[6] や小林ら[7]，藤村ら[8] のように，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をとることが考えられる．しかし，これらの手法では施設名をタイトルに含んでいるものに限り抽出することで，1 文中だけに検索対象と評価表現が含まれている場合以外の評価文も抽出できるという本研究の利点が失われてしまう．そのため，項目が入力された場合に評価値との係り受けを利用するなどの改善策が考えられる．

③ システム全体の問題点

②で述べた問題点以外について述べる．ブログ抽出の問題として，対応できるブログの割合の低さが挙げられる．つまり，現在対応しているブログサイト以外のアクティブユーザの多いブログサイトへの対応が求められる．

また，記事の見出しに施設名が 2 つ以上入っている場合，そのページにおけるすべての評価文が複数の施設に格納されてしまうという問題がある．この問題は精度に関わる重要な問題なので今後の課題として挙げられる．

参考文献


[1] 総務省．ブログ・SNS の現状分析及び将来予測．
http://www.soumu.go.jp/menu_news/s-news/2005/pdf/050517_3_1.pdf.

[2] 総務省．ブログ及び SNS の登録者数．
http://www.soumu.go.jp/menu_news/s-news/2006/060413_2.html.

[3] 総務省情報通信研究所．「ブログの実態に関する調査研究」の結果．
http://www.soumu.go.jp/iicp/chousakenkyu/data/research/survey/telecom/2008/2008-1-02-1.pdf.

[4] NEC 総研．ブログ・SNS 利用の現在～利用者アンケート調査から～．
http://www.blwisdom.com/itbz/02/.

[5] 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).

[6] 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).

[7] 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).

[8] 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).

[9] 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).

[10] Yanhong Zhai，Bing Liu：Web Data Extraction Based on Partial Tree Alignment，Proc. 14th Int'l Conf. World Wide Web，pp.76-85 (2005).


５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計１件）

① 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，情報処理学会論文誌．Vol.50，No.9，pp.2182-2192，2009 年，査読有
http://id.nii.ac.jp/1001/00066471/

〔学会発表〕（計１２件）

① 松石浩輔，鈴木雅人，松本章代，北越大輔：文章表現の癖抽出に基づく手書き文章認識の後処理方式の検討，電子情報通信学会 2012 年総合大会，2012 年 3 月 21 日，岡山大学

② 鈴木雅人，松石浩輔，北越大輔，松本章代：確率ネットワークを用いた手書き文書認識の後処理方式の検討，電子情報通信学会 2012 年総合大会，2012 年 3 月 21 日，岡山大学

③ 宇田賢広，松本章代，小西達裕，高木 朗，小山照夫，三宅芳雄，伊東幸宏：Web ペー

ジにおけるナビゲーション領域検出を利用した非主要部分特定手法,第4回データ工学と情報マネジメントに関するフォーラム DEIM, 2012年3月3日, シーサイドホテル舞子ビラ神戸
④ 清水口星也, 松本章代:歩行者を対象としたスマートフォン用ナビゲートアプリの開発,平成23年度 第6回情報処理学会東北支部研究会, 2012年2月14日, 東北学院大学
⑤ 松本章代, Martin J. Dürst：可読性の指摘を行うプログラミング教育システムの開発 ― 反復構造の自動検出による関数化の指摘 ―, 情報処理学会 第73回全国大会, 2011年3月3日, 東京工業大学
⑥ 鈴木雅人, 大久保貴博, 北越大輔, 松本章代:永字八法に基づく手書き文字認識用辞書の動的構成法, 電子情報通信学会 2011年総合大会, 2010年3月17日, 東北大学
⑦ 松本章代, 草桶慎太郎, Martin J. Dürst:プログの評判情報を用いた施設情報検索, インタラクション2010,2010年3月1日, 学術総合センター
⑧ 曽山裕, 松本章代, Martin J. Dürst：集合被覆アルゴリズムに基づいた HTML 構造内スタイルの自動 CSS 化, 電子情報通信学会 2010 年総合大会 ISS 特別企画 学生ポスターセッション,2010年3月18日, 東北大学
⑨ 押見悠平, 曽山裕, 松本章代, Martin J. Dürst：同等プロパティ抽出による CSS の最適化, 電子情報通信学会 2010 年総合大会 ISS 特別企画 学生ポスターセッション, 2010年3月18日, 東北大学
⑩ 沙鵬, 松本章代, 小西達裕, 高木朗, 小山照夫, 三宅芳雄, 伊東幸宏：繰り返し構造の検出に基づく Web ページの見出しの階層構造の解析,情報処理学会 第75回デジタルドキュメント研究会, 2010年3月5日, 沖縄県青年会館
⑪ Akiyo Matsumoto, Tatsuhiro Konish, Akira Takagi, Teruo Koyama, Yoshio Miyake, Makoto Kondo, Yukihiro Itoh: Judgment of Web Pages Based on Information Value that Decays with Time, The 4th International Conference on Ubiquitous Information Management and Communication, 2010年1月14日, Sungkyunkwan University (韓国)
⑫ 松本章代, 小西達裕, 高木朗, 小山照夫, 三宅芳雄, 伊東幸宏: 時間の経過により価値が減衰する情報を主体とするウェブページの判定, Web とデータベースに関するフォーラム 2009, 2009年11月19日, 慶應義塾大学

〔その他〕
ホームページ等
http://mmtl.cs.tohoku-gakuin.ac.jp/research/reputation.html

6．研究組織
(1)研究代表者
松本　章代 (MATSUMOTO AKIYO)
東北学院大学・教養学部・講師
研究者番号：40413752