ジオコーディング技術による釣りブログの可視化

2007MI032 原田　健太　2007MI056 堀山　洋輔
指導教員　河野　浩之

1　はじめに

近年，MAP と別情報を融合したマッシュアップサイトは多数存在する．釣りに関してのこのようなサイトは WEB 上では様々な釣り場情報が存在し，それを基に釣り場へ行くユーザも少なくないにも関わらず，商業者など一部の人が登録したものしか存在しない．また，WEB 上には個人ブログなど釣り場情報を含むものが多数存在し，そのなかには雑誌などにも載っていない穴場情報などが記載されているものも存在する．しかし，ブログにはブロガーの近況など，余計な情報が入っており，必要な情報を調べるのは時間がかかる．

本研究では，ユーザが釣り情報に関するキーワードを入力し，検索にヒットした情報を Google Map 上に表示できるようにする．Google Map 上に表示する内容は，地名，ブログ名，ブログにその情報が記載された日付，魚名，その釣り場の評価である．また，プラスの評価を表す釣り場は赤いマーカー，マイナスの評価を表す釣り場には青いマーカーをそれぞれ表示する．このシステムの実現方法として，収集した釣りブログに対して ChaSen で作成した魚名テーブルと照合し，関連語を抽出する．データはにほんブログ村で紹介されている釣りブログから収集する．関連語として抽出された地名は CSV アドレスマッチングサービスを用いて作成した地名テーブルと照合し，緯度経度に変換する．評価表現の抽出，スコアリングは ChaSen が抽出した単語をあらかじめ用意した評価語辞書と照合することにより行う．

2　ブログ解析方法の諸研究

ブログ解析技術を用いた論文を紹介していく．表 1 はジオコーディング技術を用いた研究をまとめたものである．

2.1　関連語抽出の先行研究

数原ら [1] の手法では一つの話題を入力とし，的確にイベントマイニングを行うための適切な関連語を抽出している．抽出方法として以下のように行う．

(1) ブログ記事から取得した大量の文を対象に，1 文ずつ係り受け解析を行い，格要素・格助詞・述語をひとつの組としたパターンの抽出を行う．

(2) 抽出された述語パターンについて，それぞれ同一のパターンの頻度を計算する．そして，入力された話題語が含まれている述語パターンに注目し，同一の述語を持つ述語パターンと組み合わせを行う．

ここでは，ふたつのベースライン手法と比較し高い精度が得られた．また竹市ら [2] は形態素解析ツールの ChaSen を使用している．

2.2　地名抽出の先行研究

安村ら [3] の手法は，地理的包含関係を考慮して，地名の未登録による地名の未抽出に対処している．さらに，同じ場所を示す異なる地名の存在にも対処するため，正式名称とは異なる地名を登録している．

2.3　ジオコーディング技術の先行研究

竹市ら [2] の手法は，抽出した地名を東京大学空間情報学研究センターが提供している CSV アドレスマッチングサービスを用いてあらかじめ地名テーブルを作成しておき，緯度経度に変換している．また細川ら [5] の手法は，翻訳対象地名文字列を「地名が指す場所の説明文」と拡大して，地図への文書自動配置機能を実現している．

3　釣り情報取得取得システムの構築

3.1　釣り情報取得取得システムの構築案

まず，あらかじめ収集しておいたブログを解析プログラムに通し，データベースに格納していく．データベースは MySQL を選択した．出力先の地図は, Google Map を選択した．ここでは本研究の地域情報取得システムの構築手順を (1)〜(6) で示し，図 1 に表示する．

(1) まず準備として釣りブログの記事を wget を用いて収集しておく．この際拡張子が html のものだけ収集する．収集した釣りブログのファイルは全て同一ディレクトリに保存し，Perl プログラムを読み込みやすくする．

(2)Perl のプログラムを実行して，収集したブログを解析しデータベースに格納していく．この際形態素解析，地名抽出，魚名抽出，評価表現抽出，評価語のスコアリング，抽出された地名の緯度経度の照合，全文検索を行うためのインデックス作成もこのプログラム内で行う．格納していく内容は，ファイル名，魚名，マップに乗せる地名，評価語のスコア，マップに載せる地名の緯度経度，全文検索を行うための全文インデックスである．

(3) 実際にユーザーが操作を行う動作で，PHP を用いてユーザーが検索したい用語をキーワード検査画面に入力する．

(4) キーワード検索画面から受け取ったキーワードを検索 PHP プログラムに送る．

表 1 ジオコーディング技術を用いたシステムの比較

タイトル	抽出対象	抽出データ	抽出方法	ジオコーディングの手法
ブログにおけるイベントマイニングのための適切なキーワード抽出 [1]	一般ブログ	関連語	同一パターンの頻度	なし
観光ブログの評価抽出による地域情報獲得 [2]	観光ブログ	地名，評価表現，評価	[3] の手法	あらかじめ作成した地名テーブルを参照
地図への文書自動配置機能の地域内情報発信システムへの適性評価 [5]	一般ブログ	関連語	同一パターンの頻度	非地理的特徴の活用による位置情報の関連付け

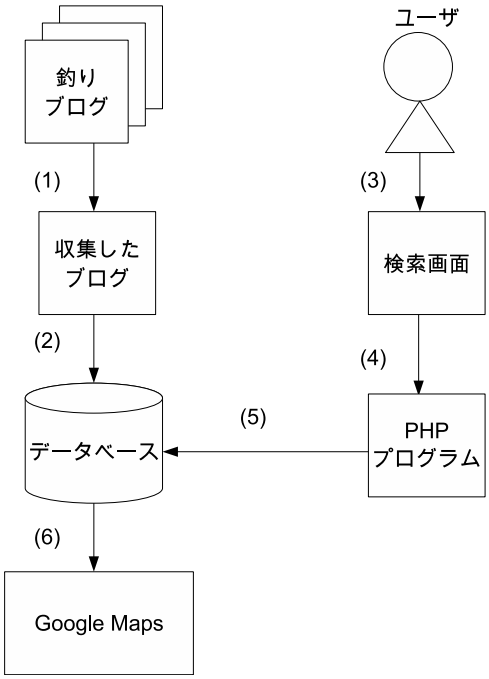

図 1　地域情報取得システムの構築

(5) 送られてきたキーワードを基に，PHP 上から SQL 文をのせてデータベースにアクセスし，キーワードが MySQL に格納されているか検索を行う．検索にヒットした情報をマップに表示できるようにする．
(6) データベースに登録してある情報から，地名，魚名，URL を記載した Google Map を作成する．

3.2　抽出した地名情報のジオコーディング

ブログからの釣り場の情報だけでは具体的な地名が得られない場合がある．そこで，安村ら [3] の手法を用いる．この手法により例として，地名情報である「名古屋市釣り場公園」が未登録であったとしても，上位階層の地名である「名古屋」を含んでいるため，「名古屋」を抽出することができる．またこの場合，ブログエントリを「名古屋市釣り場公園」に対応付けることはできないが，地理包含領域の地名「名古屋」に対応付けすることができる．また，このようにして抽出した地名に，ジオコーディング技術の CSV アドレスマッチングサービス使用し，緯度経度を付加するためには住所が必要となる．しかし釣りブログでは，具体的な住所を抽出できないため，地名を抽出し住所を付加させる必要がある．

本研究では，この問題を解決するために，あらかじめ地名テーブルを作成しておき，それを参照することで可能とする．地名テーブルの作成方法は，CSV 形式の住所データを日本郵政のサイトから市町村レベルで取得し，CSV アドレスマッチングサービスを通して地名に緯度経度を付加したテーブルを作成する．そうして抽出した地名をこのテーブルに通すことで， 緯度経度に変換していく．

4　釣り情報取得システムの実装

4.1　釣りブログデータの収集

本研究での釣りブログの収集にあたって，「にほんブログ村-釣りブログ」[*1] から収集し行う．収集するブログの記事は東海地方のブログ 100 件と限定した．また，ブログ記事をダウンロードするため，クローラ機能を持つ wget を使用し，「DL.bat」というファイルを作成した．これは「test.txt」内に記述された URL から拡張子が html となっているものを再帰的にダウンロードする．

4.2　釣りブログの地名抽出

地名抽出のプログラムを地名抽出プログラム枠内に示す．地名抽出を行うため，まず chomp ($_) によって改行を行うコードの削除を行う．読み込んだファイルは空白により単語が区切られているため，@blog = split(/\s+/, $_); によって空白で分割し，配列$blog に 1 つずつ格納していく．また$blog[$i] =~ s/(|)+//g; では全角，半角の空白を削除する．

具体的な地名抽出の方法としては上の枠内で記述されており，形態素解析で得た結果を用いて行っていく．形態素解析の結果における地名の品詞は「名詞-固有名詞-地域-一般」と分析されるため，基本的な方針としては，この文字列が出てきた時，地名の抽出を行うものとする．ただし例外として，「愛知県愛知郡」という地名が出現した場合，「愛知」という単語を 2 回地名と認識してしまい，地名の正確な出現回数が分からなくなってし

[*1] http://fishing.blogmura.com/

まう．これに対応するため， $hozon = $blog[]; を用いて，1 回前に出現した単語と形態素解析が分析した品詞を記憶し，地名の後に「県」という単語が出現した場合は地名抽出を行わないこととした．また地名の出現回数に関しては，地名が抽出されたとき$ count++; を行い，$ count の値を 1 増加させて，地名を chimei-extract.txt に書き込む．また，もし地名が抽出されなかった場合は，$ count が 0 の場合で，その場合は No_ place と chimei-extract．txt に書き込まれる．

地名抽出プログラム

```
while(<EXTR>){
    chomp ($_);
    @blog = split(/\s+/, $_);
    for($i=0; $i<@blog; $i++) {
$blog[$i] =~ s/( | )+//g; }
    for($i=1; $i = 2; $i++){
if(($hozon2=~ /名詞-固有名詞-地域-一般/)
   && ($blog[0] !~ /県/)
   && ($blog[3] !~ /名詞-一般/)
   && ($hozon1 !~ /EOS/)) {
    $count++;
        print OUT "$hozon1 ";
$hozon1 = $blog[0]; $hozon2 = $blog[3];
else{$hozon1 = $blog[0];
     $hozon2 = $blog[3];}
if(($blog[0] =~ /漁港/)
   &&($hozon4 =~ /名詞/)){$count++;
        print OUT "$hozon3$blog[0] ";}
$hozon3 = $blog[0];$hozon4 = $blog[3];
    last;}
else{$hozon3 = $blog[0];
   $hozon4 = $blog[3];
    last;}}}
close(EXTR);close(OUT);
```

4.3 評価語のスコアリング

評価語のスコアリングを行うプログラムを評価語のスコアリングプログラム枠内に示す．評価語のスコアリングについては，最初に$ward =$_; で評価語が入っているテキストファイルから評価語を読み込む．次に，open(JISYO， $in) で鍛冶ら [6] のスコア辞書を参考にした，釣り用のスコア辞書を開く．これは，釣りの評価語には「入れ食い」,「坊主」など釣りの専門用語が使われており，そのままでは釣りの評価語を抽出できないためである．while(<JISYO>) の内部で改行を削除し，空白で分割し，@data という配列に内容を格納する．評価語の内容と釣り用のスコア辞書の内容が一致した時に $score += $data[0]; とすることで，スコアの合計を計算している．また，もし評価語がなかった場合は「評価語ではありません」と出力される．

評価語のスコアリングプログラム

```
while(<HYO>){
    chomp ($_);
    $ward = $_;
    $in = "sukoajisyo. txt";
    open(JISYO， $in) or die
("can't open file $in\n");
    while(<JISYO>){
chomp ($_);
@data = split(/\s+/， $_);
for($k=0; $k<@data; $k++) {
    $data[$k] =~ s/( | )+//g;}
if($ward eq $data[1]){
    $score += $data[0];
    print OUT "$ward $data[0] \n";}
elsif($data[1] eq "EOS"){
    print OUT "評価語ではありません\n";}}
    close(JISYO);}
```

4.4 釣り情報取得の各プログラム

釣り情報取得システム構築の際に，4 つのファイルを作成した．各ファイルの説明し，図 2 で地域情報獲得システムの実行例を示す．

zen.pl: ブログを読み込み形態素解析，地名抽出，魚名抽出，評価表現抽出，スコアリング，抽出された地名の緯度経度の取得，データベースに各情報の格納を行っている．kensaku.php: ユーザーが検索したいキーワードを入力する画面．

BLOG.php: kensaku.php から受けた取ったキーワードをデータベースにアクセスし，ヒットした情報を XML データに変換し，BLOG.php に送る．

map.html: BLOG.php から受けとった XML データを基に地図作成．

5 釣り情報取得システムの考察評価

実装したシステムに対してアンケートを実施し，その結果を基に考察評価を行う．なおアンケートの比較対象として，「にほんブログ村」の中部地方，「釣りなび」[*2] の東海地域を対象とした．これは，本研究が東海地方のブログ 100 件を対象としているためである．この 3 つのシステムについて，各アンケート項目の実演を実際に見てもらい評価してもらう．アンケートは学生研究室の 10 人を対象に実施し，以下の 5 つの項目について，5 段階で評価してもらった．なお (6) は自由回答とした．

(1) ユーザの入力した情報 (地理情報，魚名) からブログの URL，地名，釣れる魚，評価をわかりやすく知ることができた．

(2) ユーザが入力した情報 (地理情報，魚名) から地名を

[*2] http://ika.s17.xrea.com/

図 2　地域情報獲得システム実行例

簡単に知ることができた.
(3) ユーザが入力した情報の評価を素早く表示できた.
(4) ユーザが検索したい項目を簡単に調べられた.
(5) システムの総合的な満足度 (操作性，情報収集能力，画面の見やすさ，検索時間の長さといった，システムの総合的な評価).
(6) 本システムの不満な点，改善点.

表 2 はそれぞれの項目のアンケート結果をまとめた物である．にほんブログ村，釣りなび内の () 内の数字は本研究とのスコアの差を表している.

表 2 各システムのアンケート結果

質問項目	本研究	にほんブログ村	釣りなび
(1)	3.8	1.6(2.2)	4.6(-0.8)
(2)	4.9	2.0(2.9)	1.8(3.1)
(3)	4.6	3.8(0.8)	3.7(0.9)
(4)	2.5	2.1(0.4)	4.7(-2.2)
(5)	3.4	2.7(0.7)	3.0(0.4)

表 2 からもわかるように，本システムは質問項目 2 と質問項目 3 では高評価だが，質問項目 1 質問項目 4 は釣りなびより低評価である．改善方法としては，地図以外にも，評価の一覧，コメントを含んだ項目の表示などのホームページの改善などがあげられる．また，システムの総合的な満足度という項目では他の 2 つのシステムより高評価を得ることができた．しかし，他の 2 つのシステムとのスコアの差はそれほどつかなかった．その理由としては，人物名などと魚名が同じ場合の対処がしていない，データベースに登録してあるブログの情報量の少なさ，キーワード検索の際の該当しないキーワードが存在する，スコア辞書の得点付けが詳細でないなどがあげられる．本システムはプロトタイプであるため，データベースに関東地方などのブログ記事の情報を多く登録していない．よって，データベースに多くのブログ記事の情報を登録し，その他の改善点を改善すれば，他の 2 つのシステムとの満足度の差は大きくなるのではないかと考えられる.

6　まとめ

本研究では，WEB 上に存在する未整理のブログからの釣り情報の収集が難しいという問題点を，ブログの形態素解析，地名と評価の抽出をし，GoogleMap 上に表示することにより解決した.

また，各項目の 5 段階評価のアンケートから本システムと他の 2 つのサイトとの性能比較を行った．ブログの URL，地名，釣れる魚，評価のわかりやすさの項目は 3.8，入力した情報から地名を簡単に知ることができたという項目は 4.9，入力した情報の評価を素早く表示できたという項目では 4.6，検索したい項目を簡単に調べられたという項目では 2.5 という結果を得た．またシステムの総合的な満足度は 3.4 であった．他の 2 つのサイトと比較して低評価だった項目は，ブログの URL，地名，釣れる魚，評価のわかりやすさに関して釣りなびが 0.8，検索したい項目の調べやすさに関して釣りなびが 2.2 上回る結果となった.

また，収集した釣りブログ 100 件の地名の抽出精度に関しては 79% という高い数値を得ることができた．釣りブログは市町村名を記述しているものが多かったため，市町村名に重点をおいた抽出方法は釣りブログに関しては実用的であると考えられる.

今後の課題としては，魚名だけでなく魚のサイズが表示できるようにすること，魚名と同じ人物名などを抽出しないようにすることがあげられる.

参考文献

[1] 数原良彦, 戸田浩之, 櫻井彰人,“ ブログにおけるイベントマイニングのための適切なキーワード抽出, ” 電子情報通信学会第 18 回データ工学ワークショップ, pp.1-6, 2007.
[2] 竹市紘一郎, 武野佑基, “ 観光ブログの評価抽出による地域情報獲得, ” 南山大学数理情報学部情報通信学科 2009 年度, 卒業論文, pp.172-175, 2009.
[3] 安村祥子, 池崎正, 渡邉豊英, 牛尼剛聡, “ blog マッピングを用いたイベント情報抽出, ”DEWS2007 D8-3, pp.1-8, 2007.
[4] 藤村滋, 豊田正史, 喜連川優,“ 電子掲示板からの評価表現および評判情報の抽出, ” 人工知能学会全国大会 (第 18 回), 3F1-03, pp.1-4, 2004.
[5] 細川宜秀, “ 地図への文書自動配置機能の地域内情報発信システムへの適性評価, ” DEIM2010 D7-5, pp.1-8, 2010.
[6] 鍛治伸裕, 喜連川優, “ 自動構築した評価文コーパスからの評価表現辞書の構築, ” 日本データベース学会 Letters Vol.6, No.1, pp.1-4, 2007.