6S-6

ブログからのお出かけ情報抽出

松本　敦志　　遠藤 拓　　杉本 徹
芝浦工業大学　工学部　情報工学科

1. はじめに

ブログは一般の Web ページに比べ、個人の感想や体験情報を多く含む傾向が見られ、それらの情報収集に有効であると考えられる。しかし、ブログは数多く存在するため、自分の目的に合った情報を見つけるのは大変な作業となる。

本研究では、Web 上のブログ記事に対し解析を行い、観光地を訪問した体験について述べた情報や感想（本稿ではこれをお出かけ情報と呼ぶ）の抽出と抽出した情報のまとめを行う。お出かけ情報としては、訪問場所の名前とその場所における感想を扱うものとする。まとめた情報を提示することで、多くのブログ記事に目を通さずともお出かけ先決定の参考になる情報を得られるようにすることを目標とする。

2. 研究内容

本研究の処理の流れを図 1 に示す。

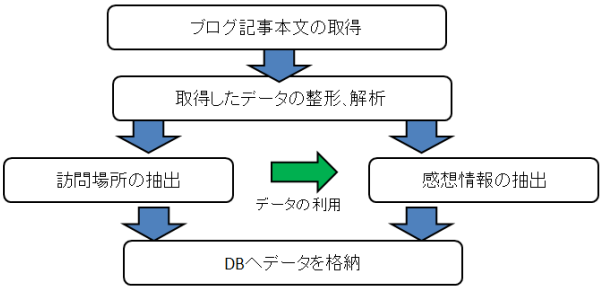

図 1:処理の流れ

2.1 ブログ記事本文の取得

Yahoo! ブログ検索 Web API を用いて適当な観光地名を含むブログ記事を検索する。検索結果に含まれる URL 情報を使用し、各記事のページソースを取得する。ページソース内に書かれた特定のタグに着目し、タグに挟まれた領域に存在する文字数、画像数からブログ記事の本文が書かれている領域を推測し抽出を行う。

2.2 取得データの整形、解析

解析作業の前準備として、取得したブログ記事本文に対して、必要のない部分の削除や特定文字列の変換といった整形作業を行う。次に形態素解析・係り受け解析を行い、文に含まれる単語の品詞や文節の係り受け関係を認識する。

2.3 訪問場所の抽出

ブログ著者が訪問した観光地名を抽出する。ブログ文中に観光地名が出現した場合でも、必ずしもブログ著者はそこを訪れたとは限らない。そのため、文中に出現した観光地名に対して、著者が実際にそこを訪問したかどうかを判断する必要がある。

2.3.1 観光地名 DB

本研究では、観光地名を集めた DB を使用している。この DB は、冨樫ら[1]が作成した観光地 DB を基として、Wikipedia から DB に存在しない観光地名を抽出し、追加することにより作成した。

2.3.2 DB 内の観光地名を含む文の抽出

ブログ文中から DB に存在する観光地名が含まれている文を抽出する。抽出した文に対し次節で提案する手法を用いることで、ブログ著者がこの観光地を訪問したかどうかを判断する。

2.3.3 観光地を訪問したかどうかの判断

前節で抽出された文について以下の調査を行い、その観光地を実際に訪問したかどうかの判断をする条件の作成を行った。

格助詞に着目をし、文中に存在する「観光地名＋格助詞」の組み合わせとその観光地に訪問しているかどうかの調査をした。その結果、格助詞「へ、に、を、から、で、にて」が観光地名の直後に出現した場合にそこへ訪問している可能性が高いと分った。

同様に動詞に着目をし、文中に出現する動詞と係り関係について調査をした。その結果、「行く、訪問する」などの動詞が存在し、それらの動詞が観光地名と係り受け関係にある場合にその観光地へ訪問している可能性が高いと分った。

さらに、お出かけ情報について記述されたブログ記事に特徴的な表現を調査した。

これらの調査の結果をふまえて最終的に以下に示す①-④の条件を作成した。

①「観光地名＋格助詞」の組み合わせが存在し、特定の格 助詞（「へ、に」など）が使用されている
②特定の動詞（「行く」など）が存在し、観光地名と動詞 が係り受け関係にある
③画像の前後に存在する文中に観光地名が出現する
④「→」や「⇒」の前後に観光地名が存在する

Extraction of Outing Information from Blogs
Atsushi Matsumoto , Taku Endo and Toru Sugimoto
Shibaura Institute of Technology

Copyright ©2011 Information Processing Society of Japan.
All Rights Reserved.

前節で抽出した文、すなわち観光地名を含む文が上記①-④の条件のいずれかを満たした場合にその文中に含まれる観光地を訪問したと判断して抽出する。

2.4 感想情報の抽出

2.4.1 パターンを用いた感想の候補の抽出

ブログの著者が観光地を訪れた記事の中で、「寺」などの名詞や観光地名を“感想の対象”、「美しい」などの名詞にかかる評価に関する単語を“感想の評価”、それらの組み合わせを“感想”と定義した。

“金閣寺”や“東京タワー”などの観光地 6 か所について、ブログ著者が訪れた記事をモデルケースとして収集し、感想に典型的なパターンや係り受け関係を定義して以下の手順で抽出を行った。

① 観光地を訪れた記事中から 1 文ずつ形態素解析と係り受け解析を行う。
② 解析した文の品詞を調べ、その並び順が下記のパターン①-③に合致する場合“感想の対象”と“感想の評価”にあたる単語を抽出する。
③ ②で抽出した単語を EDR 電子化辞書[2]で検索し、単語の概念が感想によく使われる概念の下位概念になっているかどうか確認する。感想によく使われる概念の例として“感想の対象”では“建造物”、“感想の評価”では“人の外見の美醜の値”などが挙げられる。

パターン①:〈対象名詞〉〈助詞〉?〈評価表現〉
パターン②:〈評価表現〉〈対象名詞〉
パターン③: 〈対象名詞〉〈助詞〉という文節が〈評価表現〉に係っている
ここで、〈対象名詞〉と〈評価表現〉は次のように定義される。
〈対象名詞〉 = 〈感想の対象〉〈[の|と]〈感想の対象〉〉*
〈評価表現〉 = 〈感想の評価〉〈評価の品詞ごとに異なる単語群〉
(評価の品詞が“名詞-サ変接続”なら[さ]+[れ]+[た]など)

2.4.2 概念を利用した重みづけによる絞り込み

前節で抽出した“感想の対象”と“感想の評価”をそれぞれ EDR 電子化辞書で検索し、単語の概念とその上位概念を求めて、下記の項目について重みづけを行い、定めた閾値以下の感想を除くことで絞り込みを行った。

① “感想の対象”の EDR 電子化辞書の検索結果における、感想に使われる可能性がある品詞の割合
②感想によく使われる概念がそれぞれ、モデルケースにおいて観光地の感想に出現した頻度
③単語の上位概念を求めた際に得られたすべての概念のうち、観光地に使われやすい概念が占める割合
④ “感想の対象”の上位概念と“感想の評価”の上位概念がモデルケース中で同時に出現した頻度

3. 評価実験

“金閣寺”、“銀閣寺”、“清水寺” それぞれについて書かれたブログ記事を検索し、本文が正しく抽出された 200 件を選び本研究の手法を用いて抽出実験を行った。

3.1 訪問場所の抽出

2.3 節で述べた手法を用いて抽出した結果と人手で抽出した結果を比較して評価をした結果を表 1 に示す。ここで、比較用結果として観光地名 DB に存在する観光地名が文中に出現した場合にそれらを全て抽出した時の結果も示した。

表 1：訪問場所抽出の実験結果

	適合率	再現率	F値
比較用結果	0.35	0.91	0.51
本研究の手法	0.68	0.78	0.73

本手法を用いた場合、比較用結果と比べて F 値が高い値となった。2.3.3 節で述べた個々の条件の精度を調査すると①の条件が他に比べやや低い値となっていた。また、不足している条件の調査、条件の組み合わせなどを検討していき精度向上を目指すことが今後の課題である。

3.2 感想情報の抽出

2.4 節で述べた手法を用いて感想の抽出を行った結果、再現率が 0.38、適合率が 0.37 という値となった。

本研究では感想を判別する際に品詞の並び方のパターンを利用しているが、ブログは一般に文章表現の多様性が大きく、表現のパターンを網羅的に与えることは難しい。

加えて直観的には近い意味を持つ単語の概念が EDR 辞書の概念階層上で離れた位置にあるケースがあり、これらの理由で再現率が低い値となった。

一方、品詞の並び方のパターンのみで抽出を行うと、本手法を用いた場合の 3 倍以上の数の文が抽出され、その大半は感想として不適切なものであった。このことから、概念を利用した絞り込みには一定の効果があると考えられる。

4. まとめ

本研究ではブログ記事に書かれた文章について調査を行い、ブログ著者が訪問した場所と感想の抽出手法の提案を行った。

今後の課題として、抽出条件の見直しや追加を行い抽出精度を向上することがあげられる。また、取得したデータの具体的な使用方法の検討もしていきたい。

参考文献

[1] 冨樫潤平, 杉本徹：「Wikipedia と Google Maps を利用した対話型観光地推薦システム」, 第 71 回情報処理学会全国大会, 2009
[2] 日本電子化辞書研究所: EDR 電子化辞書第二版, 2001.

Copyright ©2011 Information Processing Society of Japan.
All Rights Reserved.