ブログを利用した観光情報リンク集の自動構築

石野亜耶
広島市立大学

小林大祐
広島市立大学

難波英嗣
広島市立大学

竹澤寿幸
広島市立大学

1. はじめに

2007 年 1 月に「観光立国推進基本法」が施行され、2008 年 10 月には国土交通省の外局として観光庁が設置されるなど、日本では「観光」を 21 世紀の基幹産業と位置付け、観光を支援する多様な取り組みが積極的に推進されている。Web 上で利用可能な観光を支援する媒体としては、地方公共団体や旅行会社などが運営する観光ポータルサイトが挙げられる。しかし、観光ポータルサイトには以下のような問題点がある。

- 人手で構築されたものであり、作成に多大なコストを要する。
- レストランやホテルのサイトへのリンクは多いが、観光名所に関する歴史やニュースなど幅広いサイトへのリンクは少なく情報に偏りがある。

そこで本研究では、旅行記が記述されたブログエントリ（旅行ブログエントリ）から自動的にリンクを収集し分類することで、低コストでの観光情報リンク集を構築する。同時に、網羅性の高さや最新の観光情報リンクを素早く獲得できる点などで、既存の観光ポータルサイトよりも有用なものになることが期待される。また、ブログ著者の属性(性別、年齢、居住域など)を文体や記載内容から自動的に推定する研究が進んでいるが[1, 2, 3]、このような技術を利用することで、例えば「女性に人気のスポット」や「若い人に人気のスポット」などユーザに適した観光情報も自動的に抽出できるようになると期待できる。

本論文の構成は以下の通りである。2 節では関連研究、3 節では提案手法、4 節では実験結果について述べ、5 節で本稿をまとめる。

2. 関連研究

Web からの地域情報の自動収集に関しては、これまでにいくつかの先行研究がある。大槻ら[4]は、Web から地域の豊富な情報を提供するサイト(地域サイト)を自動収集し、地域サイト内のページを自動分類する手法を提案している。また相良ら[5]は、Web を対象とすることで、電話帳に未登録の新規店舗を発見する手法を提案している。本研究では旅行ブログエントリを情報源とすることで、観光情報に特化したリンク集の自動構築を目指す。

旅行ブログやそのエントリを登録したポータルサイトとしては、‘Travel Blog’[1]、‘旅行・観光ブログ村’[2]、‘フォートラベル’[3]などがある。これらのポータルサイトでは、ブロガーが自身のブログを旅行ブログとして登録することで、旅行ブログの集積を行う。しかし、ブログ空間にはたくさんのブログが存在するため、このようなポータルサイトに登録されていない一般ブログの中にも旅行ブログエントリが多数存在する。一般ブログに焦点を当てることで、様々な層のより多くの旅行ブログエントリを収集できると考えられる。

次に、リンクの分類に関する研究について述べる。Martineau ら[6]はブログ中のリンクについて、次の 3 つの観点から分類を試みている。実験では単語 uni-gram を素性とし、サポートベクトルマシンを用いて分類器を構築している。

- なぜ著者はリンクを張るのか？
- 著者は何を指摘しているのか？
- 著者はどのように感じているのか？

本研究では、観光情報に特化したリンク集の構築を目標としているため、リンクは 3.2 節で説明するタイプに分類する。

3. ブログを利用した観光情報リンク集の自動構築

ブログを利用した観光情報リンク集の自動構築の段階は、以下の 2 つのステップに分けられる。

Step1. 旅行ブログエントリの検出
Step2. 旅行ブログエントリからの観光情報リンク集の自動構築

この 2 つのステップについては 3.1、3.2 節で、それぞれ説明する。

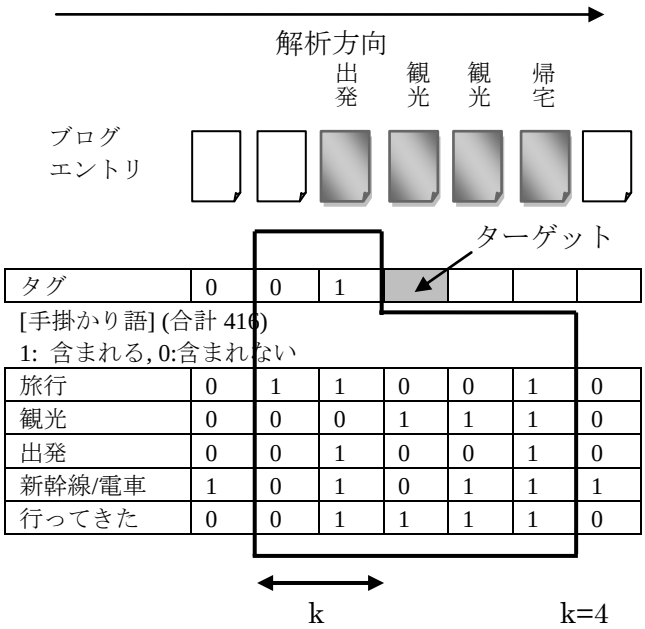
3.1 旅行ブログエントリの検出

Step1 の旅行ブログエントリの検出は、石野ら[7]の手法を用いて行う。石野らの手法について以下で説明を行う。旅行ブログエントリには‘旅行’、‘観光’、‘ツアー’などの旅行に関する手掛かり語を含む可能性が高いと言える。しかし、すべての旅行ブログに、このような手掛かり語は含まれているわけではない。例えば、あるブロガーがノルウェー旅行について複数のブログエントリにわたって日記を書いていた場合、最初のエントリには‘私たちはノルウェーに旅行に行った’と書いてあっても、2 ページ目のエントリには‘野生の羊にあったんだ！’としか書かれていないこともある。この場合、2 ページ目のエントリには旅行に関連した表現が含まれていないため、2 ページ目のエントリを旅行ブログエントリであると判定

[1] http://www.travelblog.org/
[2] http://travel.blogmura.com/
[3] http://4travel.jp/

することは困難である。そこで石野らは、それぞれのターゲットとなるエントリについてのみ見るのではなく、前後のエントリにも注目し、旅行ブログエントリの検出を系列ラベリング問題として解き、機械学習を用いて解決する手法を考案している。機械学習の手法には、CRF を使用した。CRF に与える素性とタグは以下のとおりである。

(1) ターゲットとなるエントリより前の k 個のエントリに付与されたタグ
(2) ターゲットとなるエントリの前に存在する、ターゲットからの距離が k 以内のエントリに存在する手掛かり語の有無
(3) ターゲットとなるエントリの後に存在する、ターゲットからの距離が k 以内のエントリに存在する手掛かり語の有無(図 1)



タグ	0	0	1				

[手掛かり語] (合計 416)
1: 含まれる, 0:含まれない

旅行	0	1	1	0	0	1	0
観光	0	0	0	1	1	1	0
出発	0	0	1	0	0	1	0
新幹線/電車	1	0	1	0	1	1	1
行ってきた	0	0	1	1	1	1	0

図 1：CRF に与えた素性とタグ

上記の手法により、日本語で書かれた約 1,100,000 件のエントリから、17,266 件のエントリを以下の精度で検出している。

表 1：旅行ブログエントリの検出

	精度(%)	再現率(%)
提案手法	86.7	38.1

3.2 旅行ブログエントリからの観光情報リンク集の自動構築

本節では、Step2 の旅行ブログエントリからの観光情報リンク集の自動構築について説明を行う。

3.2.1 リンク集構築の手順

リンク集構築の手順を以下で説明する。

1. Step1 で検出した旅行ブログエントリのテキストを入力する。
2. 入力テキストからリンク部分を見つけて、そのリンクに関する情報が記述されている文(引用個所)を抽出する。
3. 引用個所を用いてリンクタイプの判定を行う。
4. システムが判定したリンクタイプの結果と、人手で判定したリンクタイプの結果の比較を行う。
5. システムによるリンクタイプ判定の精度、再現率を出力する。

3.2.2 引用個所の抽出

引用個所の抽出について説明を行う。リンクに関する情報は、リンクの周辺に記述される傾向があるが、リンクから離れた場所にも記述される場合もある。よって本研究では、引用個所を手掛かり語を用いて自動で抽出する。サイトを紹介する際には、リンク先サイトのタイトルが‘「」’や‘『』’などの記号で囲まれている場合がある。また、‘紹介’、‘HP’などの語が使われるため、これらを手掛かり語として使用した。手掛かり語を以下に示す。

■手掛かり語 (26 個)
・‘「」’、‘『』’などリンク先サイトのタイトル周辺に使用される記号 (6 個)
・‘紹介’、‘HP’、‘公式サイト’、‘こちら’などリンク先サイトを紹介する際に使用される単語 (20 個)

次に、手掛かり語を用いた引用個所の抽出手順を示す。

■引用個所の抽出ルール
1. リンクが含まれている文を抽出する。
2. リンクが含まれている文の前後 X 文を抽出する。(予備実験より X=2 とする。)
3. リンク先サイトを指し示す語(Keyword)を、リンクが含まれている文、リンク前後 X 文から抽出する。
   ・手掛かり語(記号)が含まれていれば、記号で囲まれている文字列を Keyword とする。
   ・手掛かり語(単語)が含まれていれば、手掛かり語の周辺の文字列を Keyword とする。
4. Keyword が含まれている文を抽出する。

3.2.3 リンクタイプ

リンクタイプは以下のように判定する。

○ S (Spot)
旅行者が訪れた名所、施設に関する情報(歴史、生息する動物など)かどうか。

○ H (Hotel)
旅行者が宿泊したホテルや宿に関する情報かどうか。

○ R (Restaurant)
旅行者が食事をとったレストラン、食べ物、食べ物を販売するお店に関する情報かどうか。

餃子スタジアムやたこせんべいの里などは、食を売りにした観光スポットであるため、リンクタイプはSとR両方に判定される。このように各リンクは複数のタイプに判定される場合もある。

S、H、Rのいずれにも判定されないものをOとする。Oに判定されたリンクには以下のようなものがある。

・旅行に持っていくために購入したデジタルカメラのサイトへのリンク
・車を運転する際のモラルを掲載したサイトへのリンク

3.2.4 リンクタイプの判定

本研究では、機械学習によりリンクタイプの判定を行う。学習には、「引用個所に出現する各単語」、「手掛かり語の有無」を素性として与える。

リンクタイプSのリンク周辺には、観光名所の名前や、‘観光’、‘見学’、‘訪れる’など、旅行者が観光名所に訪れた際によく使われる単語が頻繁に出現すると考えられる。このような手掛かり語をWikipediaなどのWebページから収集しリストを作成した。R、Hについても同様の観点から手掛かり語の収集を行った。

○ Sの手掛かり語 (17,812個)
- Wikipedia から収集した観光名所の名前 (17,371個)
- ‘動物園’や‘博物館’など観光名所の名前に使用される単語 (138個)
- ‘見学’や‘散策’など観光の際に使用される単語（172個）
- その他 (131語)

○ Hの手掛かり語 (73個)
- ‘ホテル’や‘旅館’など宿泊施設の名前に使用される単語 (9個)
- ‘フロント’、‘客室’などの宿泊施設の構成要素 (29個)
- ‘泊る’や‘チェックイン’など宿泊する際に使用される単語 (14個)
- その他 (21個)

○ Rの手掛かり語 (3,028個)
- Wikipediaから収集した料理名 (2,779個)
- Wikipediaから収集した料理の種類 (114個)
- ‘レストラン’や‘食堂’など食事をとる施設の名前に使用される単語 (21個)
- ‘食べる’や‘おいしい’など食事をとる際に使用される単語 (52個)
- ‘ご飯’や‘料理’など、食べ物を指す単語 (31個)
- その他 (31個)

4. 実験

3節で述べた提案手法の有効性を確かめるため、実験を行った。

4.1 実験手法

■実験に用いるデータ

Step1で旅行ブログエントリとして検出されたエントリは17,266件であった。これらの旅行ブログエントリには7,421件のリンクが含まれていた。リンクの中には、Wikipediaやブログ、ニュースサイトへのリンクなど、リンク先URLからリンク先サイトを判定することができるものも含まれている。よって本研究では、そのようなリンクを除外した4,155件のリンクから、1,000件のリンクを抽出し、人手でリンクタイプの判定を行った結果を機械学習に用いる。人手でリンクタイプの判定を行った結果を表2に示す。

表2：1,000件のリンクに含まれる各タイプの件数

リンクタイプ	S	H	R	O
リンク件数	352	99	343	250

■機械学習

リンクタイプの判定の学習にはTinySVMを用いた。2次の多項式カーネルを使用し、4分割交差検定を行った。

■評価尺度

評価尺度は、以下に示す精度と再現率を用いた。

$$\text{精度} = \frac{\text{システムが検出した正解数}}{\text{システムが検出した数}}$$

$$\text{再現率} = \frac{\text{システムが検出した正解数}}{\text{人手で判定した正解の数}}$$

4.2 実験結果

提案手法による実験結果を表3に示す。

表3：実験結果

リンクタイプ	精度(%)	再現率(%)
S	72.7	62.5
H	81.3	64.9
R	76.7	71.9
O	48.6	71.6

上記の実験結果より、S、H、Rの各タイプにおいて、精度・再現率ともに高い数値を記録した。よって提案手法の有効性を示せたといえる。Oの精度が低くなってしまったのは、S、H、Rのいずれにも判定されないものをOとしたためである。S、H、Rの更なる精度の向上により、Oの精度も改善できると考えられる。

4.3 考察

本節では、次の各段階に分けて、リンクタイプの判定誤りの原因ついて考察を行う。

(1) 引用個所の抽出
(2) リンクタイプの判定

以下に、それぞれの段階について説明する。

(1) 引用個所の抽出
本研究では、まず旅行ブログエントリから引用個所を抽出し、リンクタイプの判定を行っている。そのため、引用個所の抽出に誤りがあった場合に、リンクタイプを正しく判定することができないものがあった。引用個所の抽出には、以下の 2 つの問題がある。

●問題 1　引用個所の抽出不足
本研究では人手で収集した手掛かり語を用いて、引用個所の抽出を行った。しかし、リンクに関する情報が記述されている文を、引用個所として抽出できていない場合があった。
1 つ目の原因として、リンクの紹介方法がブロガーにより大きく異なるため、人手で収集した手掛かり語では対応できなかったことが挙げられる。これは、リンク周辺に出現する単語を収集し、手掛かり語を網羅的に集めることで解決できると考えられる。
2 つ目の原因として、手掛かり語に頼った抽出手法では、文間の語彙的なつながりを見つけることが困難であることが挙げられる。これは、引用個所の抽出に、語彙的伝搬の情報を加えることで解決できると考えられる。

●問題 2　引用個所の過抽出
旅行ブログエントリにリンクが連続して出現している場合、他のリンクに関する情報を、ターゲットとしているリンクの引用個所として抽出する場合があった。また、リンクの直前や直後に、リンクに関係のない記述があるときに、その文を引用個所として抽出してしまう場合があった。このような場合は、他のリンクとの距離や、リンクの前後の文の語彙的なつながりを考慮に入れることで解決できると考えられる。

(2) リンクタイプの判定
本研究では、人手により収集した手掛かり語を用いた、リンクタイプの判定手法を提案した。リンクタイプの判定誤りの原因として、手掛かり語の不足が考えられる。例として、リンクタイプを R と判定する場合を挙げる。
リンクタイプを R と判定する際の手掛かり語として、‘おいしい’など食事をとる際に使用される単語を使用した。しかし本研究では、旅行ブログエントリを情報源として使用しているため、同じ‘おいしい’という意味でも‘おいしー’、‘おいし～’、‘美味しい’、‘オイシイ’など様々な記述が存在する。このため、人手により手掛かり語を網羅的に収集するのは困難である。この問題を解決する手法として、レストランの口コミサイトなどの口コミを利用することで、より多くの手掛かり語を収集することが考えられる。これにより、精度・再現率の更なる向上が期待できる。

5. おわりに
本研究では、旅行ブログエントリに存在するリンクを抽出し、リンクタイプを判定する手法を提案した。高い精度でリンクタイプの判定を行うことができ、提案手法の有効性が確認された。

参考文献

[1] Norihito Yasuda, Tsutomu Hirao, Jun Suzuki, and Hideki Isozaki. 2006. Identifying bloggers' residential areas. Proceedings of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, pp.231-236.

[2] Daisuke Ikeda, Hiroya Takamura, and Manabu Okumura. 2008. Semi-supervised learning for blog classification. Proceedings of the 23rd AAAI Conference on Artificial Intelligence, pp.1156-1161.

[3] Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James Pennebaker. 2006. Effects of age and gender on blogging. Proceedings of AAAI Symposium on Computational Approaches for Analyzing Weblogs, pp.199-205.

[4] 大槻 洋輔, 佐藤 理史. 2001. 地域情報ウェブディレクトリの自動編集, 情報処理学会論文誌, Vol.42, No.9, pp.2310-2318.

[5] 相良 毅, 喜連川 優. 2007. Web からの効率的な新規店舗の発見・登録支援手法. 情報処理学会論文誌, Vol.48, No.SIG_11(TOD_34), pp.49-57.

[6] Justin Martineau, Matthew Hurst. 2008. Blog Link Classification . Proceedings of International Conference on Weblogs and Social Media.

[7] 石野 亜耶, 難波 英嗣, 田熊 遥, 尾崎 貴紘, 小林 大祐, 竹澤 寿幸. 2009. ブログからの観光情報の自動抽出. 電子情報通信学会第 15 回 Web インテリジェンスとインタラクション研究会，pp.19-23.