実体験情報を含むBlogを抽出するシステム

佐々木建†　小山聡　栗原正仁
北海道大学大学院情報科学研究科

1 はじめに

Blog の爆発的な普及により，個人が Web 上に情報を発信することが容易になった．Blog 等の一個人が発信する情報は一般の Web ページの情報よりも主観的な感想，意見，体験等の評判情報が記述されている可能性が高く，それらの評判情報を求めている人にとって有益な情報源となっている．しかし，Blog の普及に伴い Web 上に蓄積される情報の量も膨大なものとなっており，Web 上から有益な評判情報を探すことはより困難になっている．この大量な知識を共有する集合から必要としている情報をうまく引き出すことが出来れば，自分自身では体験したことがなくても事前に知識を得られることなどに役立てると考えられる．

本稿では，数多くの Blog の中から本人が実際に訪れた場所についての体験が書かれている Blog を抽出する方法について述べる．場所に関する実体験を綴った文章から得られる情報は，その場所に旅行で訪れる予定がある等で関心を持っている人にとって，施設利用の方法や他の人がどの様に活用しているかなどの情報源として有効に活用されることが期待できる．

以下，2 章では関連の先行研究について述べる．3 章では提案手法の説明を行う．4 章で計算機実験による結果を示し，最後に 5 章でまとめる．

2 先行研究

本稿との関連する研究として評判情報や体験情報の分析や抽出が上げられる．また，取り扱っている文章が Blog であるため Blog から情報抽出を行っている研究を紹介する．

池田ら[1]や阿部ら[2]は商品やサービス等に対する個人の動向や経験の情報を抽出することを行っている．

倉島ら[3]や西原ら[4]は体験談を可視化できる様なインタフェースやシステムを構築して，体験談の獲得支援の方法の提案を行っている．

A System to Extract Blogs Including Real Experience Information
Takeru SASAKI, Satoshi OYAMA and Masahito KURIHARA
Graduate School of Information Science and Technology
Hokkaido University
† t sasaki@complex.eng.hokudai.ac.jp

3 提案手法

本章では，本稿において扱う体験情報についてと抽出手法の説明をする．

3.1 体験の定義

本稿における体験情報とは，Blog 著者本人が実際に場所に訪れて何らかの行動したことが記述されている文章のことである．例えば，次の 2 つの文は札幌市内における有名展望スポットである藻岩山に関する記述である．

(1) 昨日，藻岩山の展望台に登った．
(2) 今朝，藻岩山に雪が降った．

(1)は Blog 著者本人が体験したことであると考えられるが，(2)は体験でなく藻岩山に雪が降ったという事実を表しているだけである．

なお，本稿では藻岩山という様な場所ワードが予め含まれている文章を扱うとする．

3.2 体験情報抽出

本稿では Blog 記事中に現れる一文ごとに評価を行い，出現する全ての評価の積み重ねで Blog が体験情報を含むかを判定する．その一文を評価する際には 3 つのステップで段階的に選り分けていく．

まず初めの 1 ステップ目では，動作表現が含まれるかを調べる．動作表現とは動きを示す表記で，何らかの行動が起きた事が文章に現れる場合である．これを見つけるために下記の様に設計した品詞の出現パターンにマッチングするかを調べる．

格助詞，(動詞｜サ変名詞)，…，助動詞「た」

これは格助詞の後に動詞またはサ変名詞が続き，動詞やサ変名詞が入らない任意の文字の後に助動詞「た」が続くということを意味している．

この品詞パターンの作成意図は，Blog に自分自身が体験した出来事を書く際には Blog を書いている時間よりも前であるので，過去表現が用

いられる場合に多用される助動詞「た」が出現すると考えたからである．

また，サ変名詞とは語尾に「する」もしくはその変化形を接続することで動作を表すようになる名詞を指す．

次の 2 ステップ目では，動作主の特定を行う．1 ステップ目で調べた動作表現の動作したのは何であるかを特定することを目的とした．文章の文節の係り受け関係を解析することで動作表現に係る主格を特定し，本人の行動であるか否かを判定する．Blog では本人が主体となっている文章のため主格が省略されていると仮定し，特別な場合を除き省略されている場合を Blog 著者本人の行動とみなした．

最後の 3 ステップ目では，今までのステップで抽出された表現が対象の場所で行われた事であるかを評価する．つまり，抽出された表現と場所との関連度をスコアとして与える．ここでは，Blog 記事中において抽出された表現の出現する文の位置と対象の場所ワードが出現する文の位置との距離の差が小さいほど関連が強いと考え，抽出された表現から場所ワードまでの距離を逆数とした値をスコアとする．距離は，2 点間に含まれる文の数で測る．

また，Blog 中の全ての文章におけるスコアの総和を Blog スコアとする．

4 実験と結果

4.1 実験

GooBlogSearch[5]を用いて「札幌ドーム」，「モエレ沼」，「藻岩山」の 3 つのキーワードから，それぞれ 100 件ずつ収集したものを予め人手により体験 Blog であるかないかを評価した．Blog 検索エンジンを用いたのは，多くのスパム Blog を取り除いた結果を得られるからである．

体験であるか否かを判定する Blog スコアの閾値は 1 とした．これは、場所キーワードと動作表現が同じ一文中に含まれればスコアが 1 になることに基づく．また，スコアの順位に並べた場合の結果についても注目する．

ステップ 1 での形態素解析に MeCab[6]を，ステップ 2 での係り受け解析に CaboCha[7]を使用する．

4.2 結果

Table1 は人手で評価した際の体験 Blog 数とシステムが体験と判断した件数及びその中での適合率と再現率を示している．元の検索結果よりも 17 ポイント～19 ポイント程高く体験 Blog が含まれることとなった．

Table2 はスコアの高い 10 件，20 件，30 件と区切った場合の適合率である．Table1 の結果は上位 40 件程の場合であるのでこれも含めた結果から，スコアが高いと対象の場所について体験情報を含む Blog である確率が高い傾向にあることがわかる．

Table1: 実験結果

	体験 Blog 数	適合率	再現率
モエレ沼	59	77.8%(45 件)	59.3%
札幌ドーム	26	42.9%(40 件)	69.2%
藻岩山	40	58.7%(46 件)	67.5%

Table2: スコアの降順に並べた場合の適合率

	上位 10 件	上位 20 件	上位 30 件
モエレ沼	80.0%	85.0%	80.0%
札幌ドーム	60.0%	40.0%	36.7%
藻岩山	80.0%	70.0%	66.7%

5 まとめ

Blog に記述する際にみられる日本語の特徴を用いたルールと場所ワードとの距離からなるスコアを用いた方法で体験の抽出を行うシステムを構築した．

今後の課題として様々な文章形式が存在するという Blog の特徴に対応するため，機械学習等の手法を取り入れることを考えている．また，体験したことを述べている Blog を地図に表示して提示するインタフェースの実装を考えている．

参考文献

[1] 池田佳代，田邊勝義，奥田英範，奥雅博: Blog からの体験情報抽出，情報処理学会論文誌，Vol.49，No.2，Feb. 2008.

[2] 阿部修也，江口萌，隅田飛鳥，大崎梓，乾健太郎: Web からの経験情報のマイニング，人工知能学会知識ベースシステム研究会，SIG-KBS-A803，2009.

[3] 倉島健，手塚太郎，田中克己: Blog からの街の話題抽出手法の提案，DEWS2005，2C-i10

[4] 西原陽子，佐藤圭太，砂山渡: 出来事の画像表現によるブログからの体験談獲得支援，知能と情報（日本知能情報ファジィ学会誌），Vol.20，No.5，2008

[5] goo ブログ検索，http://blog.goo.ne.jp/

[6] MeCab，http://mecab.sourceforge.net/

[7] CaboCha，http://chasen.org/~taku/software/cabocha/