知能と情報（日本知能情報ファジィ学会誌）Vol.20, No.5, pp.000－000（2008）

原著論文

出来事の画像表現によるブログからの体験談獲得支援†

西原 陽子 *1・佐藤 圭太 *2・砂山 渡 *2

個人の体験談や日々のニュースに関するブログが記述されるようになってきた．ブログサイトに検索キーワードを入力し，得られるブログを読むことで個人の体験談を獲得できるようになっているが，得られる検索結果には体験談以外が記述されたブログも多数含まれる．このため，全てのブログ記事に目を通すことで，体験談獲得の効率が悪くなると予想される．

そこで本稿では画像を用いてブログから体験談を獲得するための支援システムを提案する．提案システムはブログの中に記述されている出来事を表す単語，すなわち，場所，動作の対象物，動作を表す単語を抽出し，抽出された単語を表す画像を用いて出来事を表現する．出来事を表す画像を表示することによって，ブログ中に体験談が記述されているかどうかをユーザに判定させる．実験によって，提案システムを用いることで体験談獲得の効率が上がることを確認した．

キーワード：体験談の獲得，ブログ，出来事の抽出，出来事を表す画像

1．はじめに

ブログを書く人は年々増加しているが，ブログには体験談や商品の感想などが記述されており，個人が持つ情報を得やすくなっている．例えば，北海道に観光へ行こうとする人が，観光ガイドには載っていない現地の情報を調べることも，ブログを検索することで可能になる．情報を調べる際にはブログサイトで「北海道AND観光」と検索キーワードを入力し，得られた検索結果を読み進めていくことになると思われる．

ここで得られる検索結果だが，北海道観光に行く予定，北海道からどこか他の場所へ観光に行ったことなど，北海道観光での体験談以外が記述されたブログも多く含まれる．このため全てのブログに目を通すことで，体験談獲得の効率が悪くなると予想される．

そこで本稿では画像を用いてブログから体験談を獲得するための支援システムを提案する．体験談には元となる出来事があると仮定し，提案システムは，ブログに記述されている出来事を表す複数の単語を抽出し，単語を表す画像を用いて出来事を表現する．画像の表示によって，ユーザにブログ中に体験談が記述されているかどうかを判定させる．これによって，検索結果の全てのブログを読む必要をなくし，体験談獲得の効率の向上を図る．体験談が記述されているときに出来事が記述されている可能性は高く，出来事の抽出により体験談の獲得を支援できると考えられる．後述の評価実験で用いた600件のブロク記事では，体験談が記述されているときに出来事も記述されている確率は93%であった．また，画像の表示によって，内容の理解に要する時間は短縮される［Hulbert79，Pietrucha85，八田98］ため，ブログに体験談が記述されているかどうかが，文章を読むよりも素早く判別できると考えられる．

本稿での出来事の定義は，「どこで，何を，どうした」を表す単語，すなわち，場所を表す単語，動作の対象物を表す単語，動作を表す単語の３種類とする．一般的に出来事を表すには「いつ，誰が，なぜ」という時間，人物，理由の情報も用いられる．このうち，時間に関しては，出来事の生起時間帯を推定する手法が提案されているが［野呂07］，機械学習の学習データを用意する必要があるため，本研究では扱わない．また，人物に関しては，日本語の文章は主語が省略されることが多く，人物の特定が難しいため抽出しない．同様に理由に関しても特定が難しいことから抽出しない．

2．関連研究

関連研究として，Webからの情報抽出，評判情報抽出，体験談獲得支援をまとめる．

† Personal Experience Acquisition Support from Blogs using Event-Depicting Images
Yoko NISHIHARA, Keita SATO and Wataru SUNAYAMA

*1 東京大学大学院工学系研究科
School of Engineering, The University of Tokyo
*2 広島市立大学大学院情報科学研究科
Graduate School of Information Sciences, Hiroshima City University

2.1 Webからの情報抽出

Webからの情報抽出に関する研究は多く[Chang06]，大勢の人に注目されている情報の抽出では，ブログの集合から注目されている人物，単語，文章を抽出する手法[Glance04]がある．また，テレビやブログで発信された情報や人間の検索履歴を用いて，注目されている人物や出来事を抽出する手法もある[Biglobe，Blog360，kizasi]．また，単語の頻度を時系列解析することによって，今後の注目が予想される単語を抽出する手法もある[Kleinberg02，Fujiki04]．本研究でもWebからの情報抽出を行うが，大勢の人によって注目されている情報ではなく，個人が持つ体験談の獲得を目標とする．

2.2 評判情報抽出

提案システムが扱う体験談は評判情報の１つと見なせるが，商品や映画の評判情報をWebから抽出する手法がある．例えば，評判をpositiveなもの，negativeなものに分類し，特徴を機械学習する手法がある[Dave03，Turney02]．また，学習した特徴を用いて商品の評価を表示するものもある[BlogSphere，Goo，Nifty]．他にも単語と格助詞の使われ方から，消費者の問題意識や希望を抽出する手法がある[松村07]．これらに対し提案システムでは体験談ではなく，体験談の元となった出来事を抽出する．

商品や映画と異なり，人間が体験する出来事は人によって感じ方や考え方がそれぞれ異なることから，出来事の記述に用いられる単語の種類に比べて，体験談の記述に用いられる単語の種類の方が多くなる[SHOOTI]．ブログ記事において，出来事と体験談は共起する可能性が高く，本研究では出来事の抽出により体験談の獲得を支援する．

2.3 体験談獲得支援

体験談獲得を支援する手法として，自発的な動作を表す表現と動作に関連する単語を抽出して，体験談を出力するシステムがある[池田07]．動作を表す単語，動作の対象となる単語を抽出する点で提案システムと似ているが，提案システムでは場所を表す単語も抽出し，出来事の詳細な把握を支援する．

ブログから個人の体験談を半自動で抽出し，ブログのサムネイル画像を表示するシステムがある[SHOOTI]．提案システムでも画像を表示するが，体験談の元になった出来事を表す画像を表示する点でこのシステムとは異なっている．

3．体験談獲得支援システム

提案システムの構成を図１に示す．提案システムは体験談を獲得したいテーマを表す検索キーワードを入力にとる．提案システムは入力された検索キーワードを用いて，ブログサイトから検索キーワードを含むブログ記事を取得する．取得したブログ記事を，出来事が記述されているものに絞り込む．絞り込んだブログ記事を場所ごとに切り出し，切り出した中から出来事を表す単語を抽出し，抽出した単語を表す画像を並べて画像列を作成する．最後に作成した画像列を出力する．

3.1 入力：検索キーワード

体験談を獲得したいテーマを表す検索キーワードをユーザが入力する．例えば，ハワイの観光での体験談ならば「ハワイAND観光」と，複数のキーワードをANDでつないで入力する．

3.2 ブログ記事の取得

入力された検索キーワードを用いて，ブログ記事を取得する．ブログ記事はブログサイトを指定して取得する．

3.3 ブログ記事の絞り込み

集めたブログ記事を出来事が記述されているものに絞り込む．日本語の文章で既に起こった出来事を記述する文は，「〜をした」と動詞の過去形で終わることが多い*1．そこで，動詞の過去形で終わる文が含まれるブログ記事を抽出することで，絞り込みを行う．

絞り込むためにブログ記事を形態素解析*2し，文末

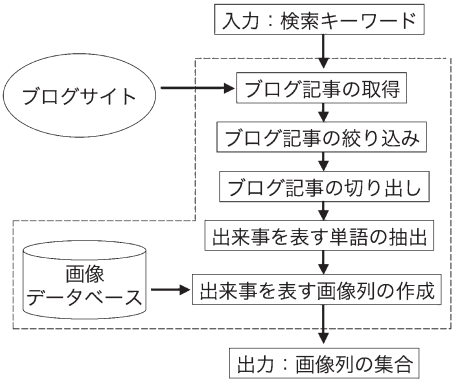

図１ 提案システム構成

*1 ブログ記事において出来事が記述された100文を調査したところ，72%が動詞の過去形で終わっていた．
*2 提案システムでは茶筌[松本97]を用いた．

に過去形の助動詞「た，だ」が含まれているブログ記事を抽出する．また，接続助詞でつなげられている文は2つに分け，後半の文が過去形の助動詞で終わっているならば，前半，後半の文それぞれを過去形の助動詞で終わっている文と見なす[ICOT]．

3.4 ブログ記事の切り分け

ブログ記事を場所ごとに切り分ける方法を説明する．1つのブログ記事には複数の場所に関する記述があると考えられる．そこで，場所を表す単語が含まれる文から，次に場所を表す単語が含まれる文の直前までを1つのブロックとし，ブログ記事を切り分ける．

次に，場所を表す単語を抽出するための前処理を説明する．場所を表す単語は名詞とし，名詞に後続する格助詞の役割を特定することによって抽出する．格助詞の役割は[渡邊05]を参考にして，表1の通りに設定する．表1を用いてブログ記事中の全ての名詞に役割の候補を与える．役割の候補が複数ある場合は，格助詞を表1に示されたフレーズに置き換え，名詞と役割の関連を式(1)によって評価する．

$$validity(p) = \frac{hit(p + phrase)}{hit(p)} \quad (1)$$

式(1)で p は場所を表す名詞，$hit(\ )$ はその名詞を含むWebページの数[*3]，$p + phrase$ は名詞の後にフレーズを付けた新しいフレーズとする．式(1)では，名詞が含まれるWebページ数に対して，新しいフレーズが含まれる割合を求めている．新しいフレーズが多くのWebページに現れるならば，名詞の役割はフレーズに相当する役割と見なせる可能性が高いと考えられる．式(1)を用いて，名詞の役割の可能性を評価し，その可能性が最も高い役割を選択することで，名詞の役割を判別する．

表1 格助詞とその役割，および式(1)の計算で用いるフレーズ

格助詞	役割	フレーズ
に, で	場所格	に行って, にて
で	道具格	を用いて
に, で, により	原因	が原因で
が, は, で	主格	の集まり
と, で, より	修飾節	のまま
を, に, と	対象格	に対して
に, へ, まで	目標格	に向かって
から, より	源泉格	から

*3 提案システムではGoogle[Google1]を用いた．

式(1)は名詞 p が現れたときに，フレーズ *phrase* が現れる条件付き確率を表している．ここでは，名詞 p が現れた時に，フレーズ *phrase* が用いられる割合を評価することが目的であるため，それを直接評価できる条件付き確率を用いている．

表1のフレーズのうち，「に行って」以外は[渡邊05]と同様のものになる．場所格に対応するフレーズとして，「にて」と「に行って」の2種類を定義したが，このうち「にて」は「車中にて」「教室にて」など特定の場所へ移動した後を表現するときに用いられることが多く，「に行って」は「山に行って」「大学に行って」など特定の場所へ移動する前を表現するときに用いられることが多い．本研究では，場所を表す多くの単語を抽出したいため，場所格のみ「にて」と「に行って」の両方を用いる．

特定された格助詞の役割を用いて場所を表す単語を抽出する．場所を表す単語は後続の格助詞の役割が「場所格，目標格，源泉格」のいずれかになるものとする．従来研究にて，場所を現す単語は主に場所格，目標格，源泉格の格助詞に接続することが確認されている[国語研97]．場所格の格助詞「に，で」は場所一般を表すものであり，目標格の格助詞「に，へ，まで」は移動先の場所を表すものであり，源泉格の格助詞「から，より」は移動元の場所を表すものであり，場所を表す単語は主にこれらの格助詞に接続し，他のものには接続しない．このことから，この3つの役割を場所を表す単語の役割とする．一文の中に場所を表す単語が複数ある場合には，式(1)の評価値が最も高い単語を採用する．

本研究では場所を表す単語を抽出する際に，格解析や固有表現，辞書などは用いないが，この理由は「学園祭で」「車中で」など地名ではない場所をこれらの手法を用いて抽出することが難しく，実際のブログ記事の中には地名ではない場所を表す単語が少なくない[*4]ためである．

3.5 出来事を表す単語の抽出

ブログ記事から出来事として場所，対象物，動作を表す3種類の単語を抽出する．3種類の単語はブログ記事が切り出されたブロックごとに抽出される．場所を表す単語は既に抽出されているので，対象物を表す単語と動作を表す単語を抽出する方法を説明する．

§1 対象物を表す単語の抽出

出対象物を表す単語を抽出する方法を説明する．対象物を表す単語は名詞とし，名詞に後続する格助詞の

*4 評価実験で用いた600件のブログ記事から被験者が抽出した単語のうち，32%の単語は格解析などでは抽出されなかった．

役割が「対象格，目標格，道具格」のいずれかになるものとする．これは，対象格は動作が作用する対象，目標格は動作の目標，道具格は「電車で山に行った」と，動作を起こさせるものを表すことが多く，対象物を表す単語には主にこれらの３つの格助詞が接続する[国語研97]ためである．一文の中に対象物を表す単語が複数ある場合には，場所を表す単語との関連を式(2)によって評価する．

$$relation(p,o) = \frac{hit(p \wedge o)}{hit(p)} \times \frac{hit(p \wedge o)}{hit(o)} \qquad (2)$$

式(2)の o は対象物を表す単語になる．式(2)は場所を表す単語が含まれるWebページ数と，対象物を表す単語が含まれるWebページ数に対して，両方の単語が共に含まれる割合を求めており，この値が高いほど，２つの単語の関連が強いことを示す．式(2)の値が最も高い単語を対象物を表す単語とする．式(2)は単語の共起に基づき，２つの単語の関連の強さを測っており，同等の結果が考えられる他の指標(カイ２乗値やJaccard係数など)によって関係を測っても良い．

対象物を表す単語はブロックごとに抽出されるため，場所を表す単語の近傍にあるものとなる．だが，ブログ記事では短い文を書く筆者が多く，場所を表す単語と同じ文には含まれないことが多いこと，そして場所を現す単語の最も近くにあると限らないことから，ブロックの中から候補を抽出し，場所を表す単語との関連を式(2)によって評価する．

§2 動作を表す単語の抽出

動作を表す単語の抽出動作を表す単語を抽出する方法を説明する．動作を表す単語は動詞とし，抽出した対象物を表す単語が含まれる文の最後に出現する動詞とする．この理由は，文章では動詞とその目的語が近くに置かれることが多く[本多82]，対象物を表す単語の近くにある動詞の方が，遠くにあるものよりも関連が強いと考えられるためである．

3.6 出来事を表す画像列の作成

抽出された単語を表す画像を用いて画像列を作成する．画像列は横に３枚の画像を並べて作るとし，左から場所，対象物，動作を表す画像とする．ブログ記事が複数に切り出されている場合は，一番初めに３種類の単語が抽出できたブロックの中の単語を用いて画像列を作成する．

画像は用意した画像データベース内にあるものを用いる．このデータベースはブログ記事から抽出された場所，対象物，動作を表す単語をYahoo!イメージ検索[Yahoo!1]で検索し，表示された上位20件の画像の中から，単語の内容を最も端的に表していると著者らが考えた画像を人手で集めることにより作られている．１枚の画像を選択することに２秒から30秒を要し，作成にはおよそ３日かかった．現在，場所を表す画像が約1000枚，対象物を表す画像が約700枚，動作を表す画像が約200枚保存されている．データベース内に画像がない場合は画像列に空白を示す．単語は評価実験に用いたブログ記事から抽出されたうち，著者らによって該当する画像が得られた単語になる．

画像列の作成では，３枚の画像は独立に選択されており，相互の依存は考慮されていない．場所，対象物，動作が表現する文脈を考慮して画像列を作成すべきだが，現状の画像検索エンジンでは「グラウンドを走った」と一部の文脈を考慮した画像検索は可能だが，「体育祭でグラウンドを走った」と全ての文脈を正確にとらえた画像検索をすることは難しい．文脈をとらえることに失敗し，表示した画像によって誤解を生むよりは，人間の知的能力である，連想能力を生かしてもらうことを念頭に，シンプルな画像を提供することを心がけた．提案システムで抽出される３つの単語は出来事，すなわち文脈を表現しており，画像の内容が元の文脈に大きく依存しない端的なものであれば，単語を画像に置き換えることで人間が文脈を予想することは可能になると考えられる．

3.7 出力：画像列の集合

作成した画像列を出力する(図２)．ブログ記事を取

図２ 提案システムの出力例：画像列付きブログ記事の集合

得する際にブログ記事のタイトルや要約[*5]も取得した場合には，画像列と合わせて出力する．

3.8 想定する提案システムの使用方法

想定する提案システムの使用方法を説明する．体験談を獲得したいユーザがいるとして，例えば図2の画面が得られた場合，ユーザは表示された画像列を参照し，出来事が記述されているブログ記事かどうかを判定する．出来事が記述されていそうなブログ記事が見つかれば，タイトルに張られているリンクから元のブログ記事を参照し，体験談が記述されたテキストを獲得する．

4．予備実験

ブログ記事の切り出しの精度を確認する実験1と，3種類の単語の抽出精度を確認する実験2を行った．

4.1 実験1：ブログ記事の切り出しの評価

提案手法によりブログ記事を切り出し，正解と比較した．

§1 実験手順

実験で用いたブログ記事はGoogleのブログ検索[Google2]で，検索キーワードを「で買った」「を食べた」「に行った」として取得したものとした．この検索キーワードにした理由は，出来事が記述されたブログ記事を取得するためである．取得したブログ記事から出来事が記述され，十分な文章量がある30件に著書らが人手で絞り込んだ．

正解の切り出し位置は次の方法で用意した．被験者にブログ記事を読んでもらい，場所が変わったと思う文を書き出してもらった．被験者は20人で，全て情報科学を専攻する大学生・大学院生の男性であった．1つのブログ記事を10人に割り当て，5人以上の被験者が書き出した文の位置を正解とした．

評価では提案システムが出力した切り出し位置と正解の切り出し位置を比較し，適合率と再現率を式(3)と式(4)を用いて求めた．

$$precision = \frac{n(system \cap participant)}{n(system)} \quad (3)$$

$$recall = \frac{n(system \cap participant)}{n(participant)} \quad (4)$$

*5 評価実験ではYahoo!ブログの検索結果で表示される要約を表示した．Yahoo!ブログでは，検索キーワードを含む前後60文字程度をブログ記事から抽出し，要約として表示していると考えられるが，他のブログサイトを用いた場合も同様の方法で要約を表示する．

式(3)と式(4)では，$n(\ )$は要素の数，$system$ はシステム出力の集合，$participant$ は正解の集合を表す．

§2 実験結果

提案システムの切り出し位置と正解の位置を比較したところ，適合率は0.75，再現率は0.49となった．再現率が低くなった原因は，ブログ記事の中でブログ作者の回想が記述される部分があり，そこで場所が変わったとした被験者が多かったためである．しかし，回想部分には出来事が記述されることは少ないと考えられ，回想部分で切り出された位置を正解の位置から除いたところ，適合率は0.69，再現率は0.77となった．このことから提案システムはブログ記事の切り出しを適切に行えると分かった．

4.2 実験2：場所，対象物，動作を表す単語抽出の評価

提案手法を用いて場所，対象物，動作を表す単語を抽出し，正解と比較した．

§1 実験手順

実験で用いたブログ記事は実験1で用いたものと同じになる．正解の単語は被験者に次の方法で用意してもらった．提案システムによって切り出されたブログ記事を示し，その中から場所を表す単語，その場所との関連が最も強い動作を表す単語，動作の対象となる単語を抽出してもらった．被験者は実験1の被験者と同じであった．1つのブログ記事を10人に見てもらい，5人以上の被験者が抽出した全ての単語を正解とした．評価では提案システムが出力した単語と正解の単語を比較し，適合率と再現率を式(3)と式(4)を用いて求めた．

§2 実験結果

ブログ記事の1つのブロックの中に含まれる3種類の単語の適合率と再現率，及びそれぞれの単語の適合率と再現率を表2に示す．1つのブロックからの3種類の単語の抽出では，適合率が0.71，再現率が0.48となった．再現率が低くなった原因は，被験者が抽出した対象物を表す単語に格助詞が後続しないものが多かったためである．だが，提案システムでは全ての出来事を抽出する必要はなく，1つのブログ記事には2つ以上の場所に関する記述が含まれていることと，場

表2 単語抽出の適合率，再現率

	3種類同時	場所	対象物	動作
適合率	0.71	0.85	0.85	0.81
再現率	0.48	0.77	0.58	0.54

所を表す単語を抽出する再現率が0.48であることから，およそ1つのブログ記事から1つの出来事の抽出が期待できる．このことから体験談獲得を支援することが可能な精度であると考えられる．したがって，提案システムは十分な精度で3種類の単語を抽出できると分かった．

5．提案システムの評価実験

ユーザが体験談を獲得する際に，提案システムは支援できるかどうかを確認する実験を行った．

5.1 実験手順

実験では，提案システムを用いて被験者にブログ記事から体験談が記述されたテキストを抽出してもらった．実験に用いたブログ記事はYahoo!ブログ[Yahoo! 2]に検索キーワードを入力し，提案システムによって絞り込まれた中から上位100件ずつとした．検索キーワードは表3に示す，観光に関する5つと，学園祭に関する1つとした．この検索キーワードにした理由は，ユーザが求める体験談を考えた場合に，未体験であるが，近い将来に体験することの事前調査が多いと考えたためである．

比較のため，ブログ記事のタイトルと要約を表示する比較システムを用意した(図3)．表示するタイトルと要約は提案システムと同じであり，提案システムと同様にもとのブログ記事を読むことができる．100件のブログ記事は25件ずつ分け，ブラウザを用いて表示した．ブラウザのウィンドウサイズは1,200×1,920とした．

実験で被験者に出した指示は次の通りであり，(2)と(3)は共通に指示した．

(1) (提案)**各ブログ記事に対する画像列を見る**
　(比較)**各ブログ記事に対する要約を読む**
(2) 各ブログ記事に検索キーワードに関する出来事が記述されているかを判定する
(3) 出来事が記述されていると思ったら元のブログ記事を読み，体験談が記述されている部分をテキストエディタ上にカットペーストする

被験者は36人で，全て情報科学を専攻する大学生・大学院生であった．1つの検索キーワードと1つのシステムにつき，18人の被験者を割り当てた．実験時間は1回につき5分とした．これは軽い調べものに費やされる時間はおよそ5分程度と考えたためである．評価では抽出された体験談数の平均を比較した．

表3 評価実験で用いた検索キーワード

沖縄, 東京, 広島, 新潟, 北海道	AND 観光
学園祭	AND 模擬店

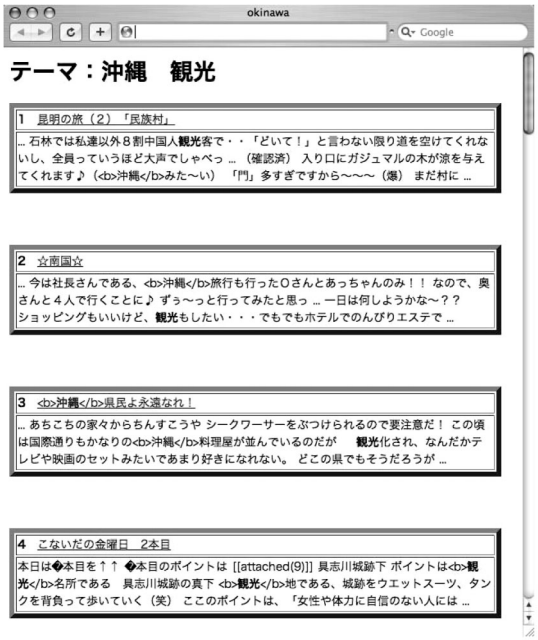

図3 実験で用いた比較システム

5.2 実験結果

被験者が読んだブログ記事数の平均を表4に示す．全ての検索キーワードで，被験者が読んだ数は提案システムを用いる方が多かった(P＜0.05)．これは表示された要約を読んで内容を理解するよりも，表示された画像を目で見て理解する方が短時間で行えたためと考えられる．このことから，より多くのブログ記事を読むことができるのは提案システムだと分かった．

次に，被験者が体験談を抽出したブログ記事数の平均を表5に示す．抽出された体験談は，体験談として妥当なものであった．全ての検索キーワードで被験者が体験談を抽出した数は提案システムを用いる方が多かった(P＜0.05)．これは提案システムでは画像に出来事が表されており，それを元に選んだブログ記事の中に体験談が記述されていることが多かったためである．このことは読んだ数に対する，体験談を抽出した数の割合を示す表6に現れており，「広島」を除き，提案システムを用いる方が高い割合で体験談が抽出され

表4 被験者が読んだブログ記事数の平均

	沖縄	東京	広島	新潟	北海道	学園祭
提案	5.8	4.9	5.1	4.7	5.9	5.4
比較	4.9	4.7	4.3	4.6	4.7	4.7

表 5 被験者が体験談を抽出したブログ記事数の平均

	沖縄	東京	広島	新潟	北海道	学園祭
提案	2.2	3.1	2.9	1.8	3.2	3.4
比較	1.4	2.6	2.6	1.3	2.5	2.1

表 6 読んだ数に対する体験談を抽出したブログ記事数の割合

	沖縄	東京	広島	新潟	北海道	学園祭
提案	0.38	0.63	0.57	0.38	0.54	0.63
比較	0.29	0.55	0.60	0.28	0.53	0.45

ていた($P<0.05$)．「広島」で比較システムの方が高くなった理由は，場所を表す画像のないブログ記事が出力されており，体験談が記述されていないブログ記事を選ぶことがあったためである．しかし，他の 5 つの検索キーワードでは提案システムを用いる方が割合が高いことから，より多くの体験談を抽出できるのは提案システムだと分かった．

以下では提案システムの方がより多くの体験談を獲得できた原因を考察する．

5.3 画像によるブログ記事の判別効果

今回用いたブログ記事の要約には「〜へ観光に行った」と出来事が記述されているものとそうでないものがあった．また，ブログ記事の中に実際に体験談が記述されているものとそうでないものがあった．この 2 点から，今回用いたブログ記事は表 7 の通りに分類される．表 7 でのブログ記事数は著者らが人手で数えた．このうち，(1)と(2)を合わせた全体の61%に相当するブログ記事は，比較システムを用いて判別することができる．だが，(3)と(4)を合わせた全体の39%に相当するブログ記事は，比較システムを用いると判別に失敗する．すなわち，(3)では体験談が記述されているのに，ブログ記事を読まないことになり，(4)では体験談が記述されていないのに，ブログ記事を読むことになる．ところが，提案システムでは(3)の場合には例えば図 4 の画像列を表示し，「学園祭でたこ焼きを食べた」ことを表す出来事が表示されていることから，要約には出来事が記述されていないが，ブログ記事の中に体験談が記述されていることを示唆できる．また，(4)の場合には例えば図 5 の画像列を表示し，検索キーワードが北海道の観光であるのに，「草津駅前の椅子に座った」ことを表す出来事が表示されていることから，目的とする体験談が記述されていないことを示唆できる．表 8 に表示された画像が出来事を表すかどうか，ブログ記事の中に体験談があるかどうかで，ブログ記事を分類した内訳を示す．画像が出来事を表すかどうかの判定は，抽出された単語が出来事を表すかどうかで行った．(3)や(4)に相当するブログ記事は表 8 の(c)と(d)になるが，(c)と(d)の合計が22.2%で，(3)と(4)の合計の39.0%を下回っていた．すなわち，被験者は画像を見ることによって，ブログ記事の判別に失敗しなかったため，より多くの体験談を獲得できたと考えられる．体験談に関わる単語を解釈する際には，頭の中で自身の経験や知識に応じ

表 7 要約中の出来事の記述とブログ記事中の体験談の有無でのブログ記事数の内訳

パターン 出来事:体験	(1) 有:有	(2) 無:無	(3) 無:有	(4) 有:無
沖縄	31	30	20	19
東京	15	60	20	5
広島	29	28	21	22
新潟	11	61	8	20
北海道	20	26	28	26
学園祭	35	20	33	12
平均	23.5	37.5	21.6	17.3
全体に対する割合		61.0%		39.0%

図 4 出来事が記述されていないが，体験談が記述されている例
(検索キーワード「学園祭 AND 模擬店」)

図 5 出来事が記述されている，体験談が記述されていない例
(検索キーワード「北海道 AND 観光」)

表 8 画像の出来事表現とブログ記事中の体験談の有無でのブログ記事数の内訳

パターン 出来事の画像:体験	(a) 有:有	(b) 無:無	(c) 無:有	(d) 有:無
沖縄	40	42	5	13
東京	28	52	7	13
広島	52	25	7	16
新潟	9	70	7	14
北海道	39	41	5	15
学園祭	41	28	4	27
平均	34.8	43.0	5.8	16.3
全体に対する割合		77.8%		22.2%

表 9 ブログ記事中に体験談が記述されていた数

学園祭	沖縄	広島	北海道	東京	新潟	平均
68	51	50	48	35	19	45.1

た，画像を伴ったイメージに置き換えた上で，連想が行われると考えられる．本研究で画像を提示することによって，その単語から画像に置き換えるというプロセスを省略できると考えられるため，単語を提示する手法に比べると，より直感的でスムーズな連想の実現が期待される．また，どの程度具体的な画像を表示すべきかについては，今後の課題であるが，シソーラスなどを用いて，画像を表す単語の抽象度別に，体験談の獲得のしやすさを測る実験を行いたいと考えている．

提案システムでは元々含まれる体験談が少なくとも，効率よく抽出できていた．体験談が記述されていたブログ記事数を表 9 に示す．「新潟」での体験談が記述されていたブログ記事数が100個中19個で，他の 5 つの検索キーワードに比べて少なかった．このため比較システムを用いると，表 6 の読んだ数に対する体験談を抽出した数の割合が0.28と最も少なくなった．だが，提案システムを用いると表 6 の割合は0.38で比較システムよりも多かった($P<0.05$)．これは，提案システムで表示される画像によって，出来事が記述されているブログ記事を判別し読むことができたためと考えられる．したがって，数が少ない場合でも画像を見て体験談の有無が判断できるため，提案システムは体験談を効率よく抽出できると分かった．

5.4 画像による出力の俯瞰効果

提案システムと比較システムでは，被験者のシステム出力の見方に違いがあった．実験終了後にアンケートをとったところ，提案システムを用いた被験者は36人中22人が「全体を俯瞰する」，4 人が「逐次見る」，10人は「俯瞰と逐次の両方」と回答した．比較システムを用いた被験者は全員が「逐次見る」と回答した．提案システムでは「全体を俯瞰する」とした人が最も多かった．この理由は画像は一瞥することで内容を把握できるため，全体を俯瞰して，体験談が記述されていそうなブログ記事を効率よく探そうとしたためと考えられる．対して，比較システムでは一瞥するだけでは内容を理解することは難しく，そのため被験者は逐次見る方法をとったと考えられる．したがって，画像を表示することでユーザはシステム出力を全体を俯瞰する傾向があると分かった．

提案システムでは全ての画像を表示できたわけではなく，表示できなかった画像は場所が17種類28件，対象物が12種類33件，動作が 9 種類40件となり，1つでも画像が表示されなかったブログ記事は600件中81件あった．表示できない画像があるならば，画像ではなく抽出された単語だけを表示することも考えられるが，これは次の 2 つの理由から体験談獲得を支援する効果が低くなると考えられる．1 つは単語を表示しては検索結果を俯瞰できず，体験談の有無を直感的に判断できなくなることがある．ブログ記事中での体験談の有無を判定させるだけならば，単語の表示によって画像と同様の効果が得ることができると考えられるが，ブログ記事を逐次調べることになり，体験談獲得の効率が悪くなる．もう 1 つの理由は，単語は画像よりもその内容の理解に要する時間が長くなることがあり[Hulbert79，Pietrucha85，八田98]，画像ではなく，単語を表示することで体験談を獲得することにより多くの時間がかかると考えられる．画像を用いる上では，端的に内容が表されていないと理解に時間がかかる欠点があり，被験者の意見にも「画像の判断に時間がかかった」とあったが，これは一部の画像に対する意見であり，全ての画像に対する意見ではない．以上，欠点はあるが，それ以上の利点の方が大きいため，単語ではなく，単語の内容を表す画像を表示する方が体験談獲得を効果的に支援できると考えられる．

5.5 場所を表す画像の効果

提案システムを用いた被験者にどの画像(画像の組合せ)に注目したかのアンケートをとったところ，表10の回答が得られ，場所を表す画像を見る被験者が多かった．出来事が記述されているかどうかを判断する時に，動作や対象物を表す単語は出来事に関する事前知識がないと判断ができない．それに対して場所を表す単語は山や海などが表示されていれば，体験談が記述されていそうだと判断がしやすい．したがって，被

表10 注目した画像の組合せと被験者の人数

場所	対象物	動作	人数
○	×	×	19
○	×	○	6
○	○	×	4
○	○	○	4
×	×	○	2
×	○	○	1
×	○	×	0

験者は場所を表す画像に注目したと考えられる．このことから，場所を表す画像が体験談獲得の中で最も効果が高いことが分かった．

表10では，場所を表す画像に注目した被験者が19人と最も多かったが，その他の画像に注目した被験者も17人居た．このことから，場所を表す画像だけでなく，その他の画像も体験談獲得を支援したと考えられる．このことから，場所を表す画像の効果が最も大きいが，その他の画像も体験談獲得の中で効果を発揮したと分かった．

6．まとめ

本稿では，出来事を表す画像列を表示することで，ブログ記事から体験談の獲得を支援するシステムを提案した．提案システムでは出来事を場所，対象物，動作を表す3枚の画像を用いて表し，ブログ記事中に体験談が記述されているかどうかを人間に判定させるための支援を行う．実験によって，画像を表示しないシステムに比べて，提案システムは同じ時間でより多くの体験談を獲得できることを確認した．

今後の課題を示す．アンケートで得られた被験者の意見に1つの画像に色々な物体が描かれていると，その内容を瞬時に判断できなかったとあった．画像の判断に時間がかかると体験談獲得の効率が悪くなるため，画像データベース内の画像を内容がより簡単なものへと交換するか，抽出された単語も合わせて表示するハイブリッド方式を検討する．そして，より多様なブログ記事を扱える汎用的なシステムにするために，どの程度まで抽象度の高い画像を用いられるかを調べるとともに，許される範囲で内容が単純で抽象度の高い画像データベースの自動構築の方法を検討していきたい．また，ブログ記事に文章の長いもの，文章の意味が分からないものがあり，体験談が記述されているテキストを探すことに時間がかかったと意見があった．今後，ブログ記事から抽出された単語を色付け表示することによって，より使いやすいシステムへと改良していく．

参 考 文 献

［Biglobe］ http://search.biglobe.ne.jp/ranking/

［Blog360］ http://blog360.jp/

［BlogSphere］ http://blogsphere.biz/

［Chang 06］ C. H. Chang, M. Kayed, M. R. Girgis, and K. Shaalan：A Survey of Web Information Extraction Systems, IEEE Transactions on Knowledge and Data Engineering, Vol.18, No.10, pp.1411－1428, 2006.

［Dave 03］ K. Dave, S. Lawrence, and D. M. Pennock：Mining the Peanut Gallery：Opinion Extraction and Semantic Classification of Product Reviews, in Proc. of the 12th International World Wide Web Conference, pp.519－528, 2003.

［Fujiki 04］ T. Fujiki, T. Nanno, Y. Suzuki, and M. Okumura：Identiication of Bursts in a Document Stream, in Workshop on Knowledge Discovery in Data Streams, 2004.

［Glance 04］ N. S. Glance, M. Hurst, and T. Tomokiyo：BlogPulse：Automated Trend Discovery for Weblogs, In WWW2004 Workshop on the Weblogging Ecosystem, 2004.

［Goo］ http://opinion.labs.goo.ne.jp/cgi-bin/index.cgi

［Google1］ http://www.google.co.jp/

［Google2］ http://blogsearch.google.co.jp/

［八田 98］ 八田一利，善如寺仁寿：プラント手順書におけるアイコンの有効性に関する研究，日本プラント・ヒューマンファクター学会誌，Vol.3，No.2，pp.127－136，1998.

［本多 82］ 本多勝一：日本語の作文技術，朝日新聞社出版局，1982.

［Hulbert 79］ S. Hulbert, J. Beers, and P. Fowler：Motorists' Understanding of Traffic Control Devices, AAA Foundation for Traffic Safety, 1979.

［ICOT］ 第五世代コンピュータプロジェクトアーカイブス，http://www.icot.or.jp/

［池田 07］ 池田佳代，田邊勝義，奥田英範：体験表現を手がかりにしたBlogの体験情報の抽出，第18回データ工学ワークショップ，A8－1，2007.

［kizasi］ http://kizasi.jp/

［Kleinberg 02］ J. Kleinberg：Bursty and hierarchical structure in streams, In Proc. of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.1－25, 2002.

［国語研 97］ 国立国語研究所：日本語における表層格と深層格の対応関係，三省堂，1997.

［松本 97］ 松本裕治，北内啓，山下達雄，平野善隆，今一修，今村友明：日本語形態素解析システム「茶筌」version1.0 使用説明書，NAIST Technical Report，NAIST－IS－TR97007，1997.

［松村 07］ 松村真宏，河原大輔，岡本雅史，黒橋禎夫，西田豊明：メッセージの背後に潜む「問い」の抽出，人工知能学会論文誌，Vol.22，No.1，pp.93－102，2007.

［Nifty］ http://shopping.nifty.com/

［野呂 07］ 野呂太一，乾孝司，高村大也，奥村学：テキスト中のイベントの生起時間帯判定，情報処理学会論文誌，Vol.48，No.10，pp.3405－3414，2007.

［Pietrucha 85］ M. Pietrucha and R. Knoblauch：Motorists' Comprehension of Regulatory, Warning and Symbol Signs, Vol.2, Technical Report Contract DTFH61－83－

C-00136, FHWA, U.S. Department of Transportation, 1985.
［SHOOTI］ http://shooti.jp/
［Turney 02］ P. D. Turney：Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews, in Proc. of the 40th Annual Meeting of the Association for Computational Linguistics, pp.417－424, 2002.
［渡邊 05］ 渡邊拓也，大野成義，太田学，片山薫，石川博：差異に注目した複数文書融合手法，DEWS2005，2C-14，2005.
［Yahoo!1］ http://search.yahoo.co.jp/images
［Yahoo!2］ http://blogs.yahoo.co.jp/

（2008年 2 月28日　受付）
（2008年 6 月30日　採録）

［問い合わせ先］
〒113-8656　東京都文京区本郷 7－3－1
東京大学大学院工学系研究科
西原　陽子
TEL：03-5841-2908
FAX：03-5841-2908
E-mail： nishihara@sys.t.u-tokyo.ac.jp

著 者 紹 介

西原（にしはら）　陽子（ようこ）［非会員］

2003阪大・基礎工卒．2005同大大学院基礎工学研究科博士前期課程了．2007同研究科博士後期課程了．2005日本学術振興会特別研究員．2008より東京大学大学院工学系研究科助教，現在に至る．博士（工学）．ヒューマンコミュニケーション支援研究に従事．人工知能学会，電子情報通信学会，情報処理学会等各会員．

砂山（すなやま）　渡（わたる）［非会員］

1995阪大・基礎工・制御卒．1997同大学大学院博士前期課程了．1999同大学院博士後期課程中退．同年同大学大学院助手．2003広島市立大学情報科学部助教授，現在に至る．博士（工学）．人間の創造活動を支援する研究に興味をもつ．言語処理学会，人工知能学会等各会員．

佐藤（さとう）　圭太（けいた）［非会員］

2006広島市立大学情報科学部卒．同年より，同大学情報科学研究科，現在に至る．

Personal Experience Acquisition Support from Blogs using Event - Depicting Images

by

Yoko NISHIHARA, Keita SATO and Wataru SUNAYAMA

Abstract：

Internet users write blogs related to their personal experience, daily news, and so on. We can obtain blogs about personal experience using search engines on the Web. However, the search engines also output blogs about other topics unrelated to personal experience. Therefore, it is necessary for us to read all blogs to obtain those about personal experiences. It takes too much time.

This paper proposes a support system for obtaining blogs about personal experiences efficiently. The system extracts three keywords that denote place, object, and action from a blog. The three keywords describe an event that leads a person to write a blog about personal experience. The system expresses the event with three pictures related to the extracted keywords. The pictures help users to judge whether personal experience is written about in the blog. We experimented with the system, and verified that it supports users in obtaining personal experiences efficiently.

Keywords：personal experience, blog, event extraction, event - depicting images

Contact Address：**Yoko NISHIHARA**
School of Engineering, The University of Tokyo
7 - 3 - 1, Hongo, Bunkyo - ku, Tokyo 113 - 8656, JAPAN
TEL ：03 - 5841 - 2908
FAX ：03 - 5841 - 2908
E - mail ：nishihara@sys.t.u-tokyo.ac.jp