Blogからの体験情報抽出

池田佳代[†1,*1] 田邊勝義[†1]
奥田英範[†1] 奥雅博[†1]

Blogに代表されるユーザが書き込む情報を参考にして，商品やサービスに興味を持ったり，購入・利用したりするユーザが増加してきている．このような情報においては，書き手の体験に基づく情報がより重視されると考えられる．そこで本論文では，ユーザの体験を記述する際に現れる特徴を体験表現として定義し，体験が記述された体験情報を提供する“体験情報抽出システム”を提案し，評価実験を行った．本実験では体験した情報が重要視されるような商品やレストランのキーワードを含むBlog記事を解析した結果，通常のキーワード検索と比較し，提案システムでは適合率が約25%高く体験情報を抽出できることを確認した．また，注目するキーワード周辺において，そのキーワードと体験表現の共起関係を調べることが効果的であることを確認した．

A Web-mining Technique of Experience Information

Kayo Ikeda,[†1,*1] Katsuyoshi Tanabe,[†1] Hidenori Okuda[†1]
and Masahiro Oku[†1]

Users who are interested in a commodity or service often turn to blogs to catch the comments made by people who have experienced the commodity or service. We develop a comprehensive set of “Experience Expressions” and use it to create a service that can extract experience information from web sites such as blogs. Experiments conducted on about 600 actual blogs (related to consumer electronics and restaurants) indicate that the proposed method is 25% more effective in extracting experience information than conventional search engines.

1. はじめに

ホテルや旅行，催し物，電化製品など様々な物事に関して，多くの人がインターネット上で個人の意見を公開している．こういったユーザの声は，評判情報としてマーケティングや商品開発，企業のリスク分析，商品購入の検討などに利用価値が高い[1),2)]．意見の公開場所は，ショッピングサイトやホテルなどの企業が運営する掲示板や個人のホームページ，Blogなどがある．特にBlogは，著者の独自の視点で記述されており，そのユーザ数も，日を追うごとに増加している．このようなBlog記事から評判情報や意見の抽出，話題語による内容の分類などの試みが行われるようになってきた[3)–5)]．

このような評判情報や意見について，「利用体験者の評判・お勧め情報が，ショップのお勧めよりも有効である」という結果が出ている[6)]．ユーザ（書き手）の体験が記述された情報は，評判情報と組み合わせたときに，文書の書き手が評価を下した理由，もしくは背景にある経験を読み手に伝えることができるため，読み手の理解が深まると考えられる．

以上の点から，ユーザの体験から得られる評判情報を抽出することができれば，消費者・生産者にとって商品の購入や選定の検討支援，マーケティングに有効である．

最近では，評判情報の検索に関する研究が試みられている[7)]が，その評判情報がどのような体験に基づいているかは考慮されていない．一方，所望のキーワードに関係するユーザの体験が記述された情報は，キーワード検索を行うだけでは，キーワードとその体験を表すような言葉の関係性を考慮して検索することはできず，単なるキーワードマッチングによる検索となり，探すことが困難である．

本論文では，分野に依存せずにユーザの体験が記述された情報を抽出する試みの1つとして，体験から得た情報を記述する際の特徴として現れる形態素の組合

†1 日本電信電話株式会社NTTサイバーソリューション研究所
NTT Cyber Solutions Laboratories, NTT Corporation
*1 現在，NTTコミュニケーションズ株式会社
Presently with NTT Communications Corporation

せを体験表現とし，その体験表現を手がかりにした体験情報抽出システムの提案とその評価について報告する．以降，2 章では関連研究を紹介し，3 章では本論文で扱う体験情報について定義し，4 章では提案する体験情報抽出システムについて紹介し，5 章では本提案システムを用いた体験情報の評価実験について説明し，6 章でその実験結果と考察，7 章でまとめを述べる．

2. 関 連 研 究

評判情報の分析としては，物事の良し悪しを抽出する際の特徴として，「良い」「悪い」といった評価内容を表す評価表現に着目し，評価表現を統計的に抽出し，それを用いて文書の肯定・否定を分類する研究がある[8]．また，“excellent” や “poor” との共起関係に関する統計量を用いて肯定・否定文を判断する研究[9]や，肯定・否定の代わりに長所・短所に分けて整理されているレポートを対象に，その製品の良し悪しに関する評価軸を自動取得する手法[10] なども研究されている．これらは，評価表現の辞書を用いずに対象物の評価を分析する手法である．

また，評価表現の辞書を用いて対象物の評価を分析する手法としては，次のような研究がある．評価表現の辞書を利用する場合，評価表現の抽出精度を向上させるためには，評価したい対象によって辞書を構築する必要がある．この特徴から，評価対象のドメインごとに評価表現辞書を作成し，意見抽出を行う手法がある[3]．しかしながら，ドメインごとに評価表現辞書を人手で作成することはコストが高くなるため，半自動的に作成する研究が行われている[11]．また，ドメインに特化した機械学習により，評価表現辞書に存在しない表現が出現した場合でも，それが評価表現かどうかを自動的に判定する研究も行われている[12]．以上のように，ドメインごとの評価表現辞書の利用や，ドメインに特化した機械学習を利用し，評価表現の抽出を行うことで，詳細な情報を取得することができる．

しかし，従来の評判情報に関する研究では，評価表現やその対象を抽出することに焦点を当てているため，宣伝広告のようなものや Blog 特有の記事自体にあまり意味を持たないものなども抽出してしまう可能性が高い．

ユーザの体験が記述された情報に関係した研究としては，Blog 記事から「地名・時間・体験」の 3 つの要素の相関関係に基づいて，観光情報に関する体験情報を抽出する試みがある[13]．

一方，池田らは分野に依存せず，体験表現と評価表現を組み合わせることで体験的評判情報を判定する試みを行っている[14]．評価表現に加え体験表現を用いることで，評価実験に用いたデータでは，宣伝のような文書を除去する効果を確認した．

本論文では，今まで未着手であった “ユーザが注目するキーワードに対する書き手の体験” が記述された情報を抽出することに焦点を当てる．

3. 体験情報の定義

本章では，本論文で用いる体験表現と体験情報について説明する．

3.1 体験表現について

ユーザの体験が記述された情報を抽出する一方法として，本論文では，体験を記述する際に現れる表現に着目する．体験を表す表現は多々存在することが想定されるが，本論文では，主に自発的な動作を表す動詞の過去形・進行形，動作を表す名詞，書き手の体験の結果として得られたであろう感想を表す形容詞の過去形などを体験表現と定義する．ただし，自己の体験ではなく，他の状態を表す「死ぬ，消える，終わる」などは含まない．また，自己の試みを表す「〜してみる」，経験そのものを表す「〜したことがある」のような言い回しも含む．**表 1** に体験表現の 5 つのタイプとその例を示す．

3.2 体験情報について

ユーザの体験が記述された情報の中で，前節で述べた “体験を記述する際に現れる体験表現” が含まれる情報を “体験情報” と呼ぶこととする．**表 2** に体験情報の例を示す．

例文 (a) は，表 1 の ① 自己の試みを表す「行ってみました」，例文 (b) は，② 経験そのものを表す「行ったことがあります」，例文 (c) は，③ 書き手が行動したことを表す「挑戦しました」，例文 (d) は，④ 行為を継続中であることを表す「通っています」，例文 (e) は，⑤ 書き手の経験から得た感想を表す「悪かった」

表 1 体験表現のタイプ
Table 1 The type of experience expression.

	表現タイプの説明	表現の例
①	「〜してみる」などの自己の試みを表す表現	行ってみた，読んでみた，試してみた，…
②	「〜したことがある」という経験そのものを表す表現	見たことがある，会ったことがある，泊まったことがある，…
③	動詞（名詞の動作を表す単語を含む）の中でも書き手自身が行動したことを表す表現	行った，食べた，使った，挑戦した，留学した，体験した，…
④	動詞（名詞の動作を表す単語を含む）の中でも，書き手自身が行為を継続中であることを表す表現	使っている，食べている，通っていた，…
⑤	形容詞（名詞の動作を表す単語を含む）の中でも書き手の経験から得た感想を表す表現	美味しかった，楽しかった，使いにくかった，…

表 2 本論文で扱う体験情報の例
Table 2 Example of experience information.

	例文	判定
(a)	最近，渋谷に新しく出来た△レストランへ*行ってみました.*	○
(b)	パリには一度だけ*行ったことがあります.*	○
(c)	岩手県でわんこそばに*挑戦しました.*	○
(d)	一月に一度は，整体に*通っています.*	○
(e)	○×ホテルのスタッフの対応は，とても*悪かった.*	○
(f)	あのマシンは安い.	×
(g)	あのホテルはアメニティと料理ではお得です.	×

のような体験表現を利用して記述しているため体験情報である．一方，例文 (f)，(g) は，表 1 に該当する体験表現を利用していないため体験情報ではない．なお，体験情報は表 2 の例にあるような 1 文だけでなく，複数文でも表現される．

4. 体験情報抽出システム

3 章で説明した体験情報をテキストから抽出するための体験情報抽出システムを提案する．

4.1 体験情報抽出システムの処理

体験情報抽出システムは，**図 1** に示すとおり，ユーザが注目するキーワード（注目キーワード）を入力すると，文書データベース（DB）にある解析対象文書を注目キーワードについての体験らしさ（スコア）でソートし，提示するシステムである．

体験情報抽出システムは，事前処理とランタイム処理に分かれている．事前処理では，解析対象文書を形態素解析し，体験表現を抽出する．ランタイム処理では，システムユーザによって注目キーワードが入力されると，解析対象文書の中での注目キーワードと体験表現との関係（スコア）を算出し，解析対象文書をランキング表示する．以降，体験情報抽出システムの詳細動作について述べる．

4.2 体験情報抽出システムの処理

まず，体験情報抽出システムは，解析対象文書の中から体験表現を抽出する．本システムでは，体験表現をルールで規定したものを利用する．この体験表現ルールは，表 1 の ①，② に示した「〜してみた，〜したことがある」というような言い回しや，③ 動詞の過去形，④ 動詞の進行形，⑤ 形容詞の過去形などを実際に形態素解析し，ルールへ変換したものを利用する．具体的な体験表現ルールの例を**表 3** に示す．体験表現ルールは品詞の組合せで設定しており，表 3 には代表的なルールを示した．その他のルールについては，表 3 のルールの応用であり，言い回しなどの語尾の細かな変化に対応したものである．

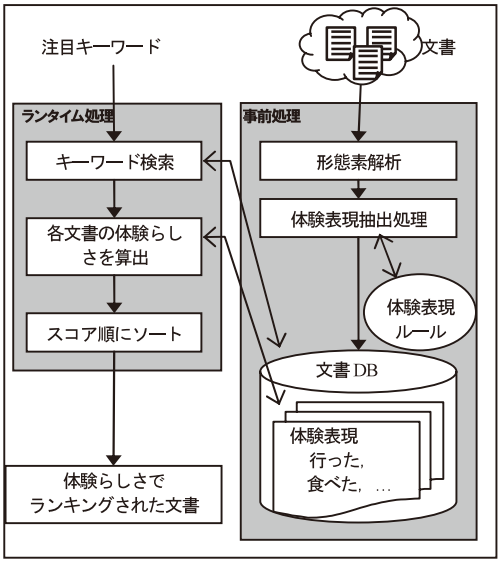

図 1 体験情報抽出システムの構成
Fig. 1 The architecture of the system.

表 3 体験表現のルールの例
Table 3 Rule of each type of experience expression.

	表現タイプの説明	表現の例
①	「〜してみる」などの自己の試みを表す表現	{動詞＋接続助詞「て／で」＋補助動詞「みた」}，{動作を表す名詞＋動詞「する（し）」＋接続助詞「て」＋補助動詞「みた」}　等
②	「〜したことがある」という経験そのものを表す表現	{動詞＋名詞「こと」＋格助詞「が」＋動詞「ある／あった」}，{①＋名詞「こと」＋格助詞「が」＋動詞「ある／あった」}　等
③	動詞の中でも書き手自身が行動したことを表す表現	{動詞＋動詞接尾辞終止「た」}，{名詞:動作＋判定詞:接続／終止「だった」}　等
④	動詞（名詞の動作を表す単語を含む）の中でも，書き手自身が行為を継続中であることを表す表現	{動詞＋接続助詞「て／で」＋補助動詞「いる／いた」}　等
⑤	形容詞（名詞の動作を表す単語を含む）の中でも書き手の経験から得た感想を表す表現	{形容詞＋形容詞接尾辞：終止「かった」}，{名詞:形容＋判定詞:接続／終止「だった」}　等

形態素解析された解析対象文書において，体験表現ルールに適合する形態素の並びが存在した場合は，体験表現があると見なす．

本ルールは，先に紹介した池田らの研究[14] により，キーワード「901iS」（携帯電話の機種名）を利用して Blog 検索したデータを用いて評価した結果，人手でラベル付けした「何らかの体験が記述されている Blog

記事（エントリ）」と「体験が記述されていない Blog 記事」を，適合率 87.4%，再現率 93.8%という高い精度で判別できている．その際の実験では，Blog 記事中に体験表現が 1 つでも存在すれば，「体験が記述されている記事」と判定した．本論文では，注目キーワードに関係する体験を抽出するために，この体験表現ルールに加えて次節に述べる 3 つのパラメータを用いる．

4.3 注目キーワードと体験情報の関係

システムユーザによって，注目キーワードが入力された後，体験情報抽出システムは，解析対象文書を基に注目キーワードと体験表現の関係性を算出する．この関係性は，注目キーワードに対する体験らしさともいえるスコアである．スコアの算出式は，次のようなパラメータを用いる．以降では，T を体験表現とし，T_i は体験表現のある種類 1 つを示し，T_{ij} は，テキスト中に出現する体験表現各々を指す．m は，体験表現の種類の総数，k は，解析対象文書の算出領域内に出現した体験表現の総数を指す．

I．算出領域：X，Y

注目キーワードの周辺に注目キーワードに関係する体験表現が出現すると考え，注目キーワード周辺テキストをスコア算出の対象領域に指定する．注目キーワードより前のテキスト領域を X，後ろのテキスト領域を Y とする．

II．注目キーワードと体験表現の共起出現頻度：$II(T_i)$

I で設定した算出領域において，出現する体験表現の中で，他の解析対象文書との間で共起数が多い体験表現は，注目キーワードとの関係性が強いことが考えられる．たとえば，「レストラン」が注目キーワードであった場合，解析対象文書の中で「レストラン」が出現するテキスト周辺において，「食べた」や「行った」という体験表現は，「泊まった」という体験表現よりも，他の解析対象文書との間で共起数が多くなることが考えられる．よって，注目キーワードを含む解析対象文書すべてにおいて，注目キーワードと体験表現の共起数が多い体験表現が出現する解析対象文書ほどスコアが高くなるような式を用いる．式 (1) は，ある種類 i の体験表現 T_i の他文書との共起数 T_i_df が注目キーワードを含む解析対象文書全体の数 N の中で出現する割合とする．このとき i は，解析文書全体において出現する体験表現の種類の数を指し，たとえば「T_1 = 食べた」「T_2 = 食べてみた」「T_3 = 行った」となる．

$$II(T_i) = \frac{T_i\_df}{N} \tag{1}$$

III．注目キーワードと体験表現の距離：$III(T_{ij})$

注目キーワードに関係のある体験表現は，テキスト中でも注目キーワードに近い位置で出現することが考えられる．よって，注目キーワードから体験表現が出現するまでの距離が近いほどスコアが高くなるような式を用いる．

上記を実現する手法として，構文解析を行い，注目キーワードとの係り受けによって，体験表現と注目キーワードの関係性を判定する方法もあるが，本論文では，解析対象文書を個人が記述する Blog のような文書とすることから，構文解析を行うコスト・誤解析のリスクを考慮して，単純な距離での判定とした．

本論文では，T_{ij}_dis を注目キーワードとある体験表現 T_{ij} との距離とし，その逆数の式 (2) を利用することとする．このとき T_{ij} とは，ある種類の体験表現 T_i が算出領域内で j 番目に出現していることを示す．たとえば「T_1 = 食べた」であるとき，その「食べた」という体験表現が，算出領域内で 3 つ出現している場合の 2 番目に出現している「食べた」を示す場合，「T_{12}」となる．

$$III(T_{ij}) = \frac{1}{T_{ij}\_dis} \tag{2}$$

以上のようなパラメータを用いてスコア算出式 $S(T)$ を式 (3) のように定義する．

$$\begin{aligned} &if(X \leq T_{ij} \leq Y) : S(T) \\ &= \sum_{i=1}^{m} \left(II(T_i) * \sum_{j=1}^{k} III(T_{ij}) \right) \\ &= \sum_{i=1}^{m} \left(\frac{T_i\_df}{N} * \sum_{j=1}^{k} \frac{1}{T_{ij}\_dis} \right) \end{aligned} \tag{3}$$

k は，解析対象文書の算出領域内に出現した体験表現の総数を指すため，T_{ij}_dis が 0 になる場合がある．このとき，式 (2) の $III(T_{ij})$ は 0 とする．このスコア算出式 (3) を利用することで，注目キーワードよりも前 X から後ろ Y という領域内で，注目キーワードの近くに出現する体験表現を持つ解析対象文書ほどスコアが高くなり，また算出領域内において，他の解析対象文書と共起数の多い体験表現を持つ解析対象文書ほどスコアが高くなる．

5. 体験表現を手がかりにした体験情報抽出の評価実験

4 章で示した体験情報抽出システムを用いて，体験情報をどの程度抽出することができるか，また，スコア（体験らしさの強度）順に並べられたときのランキングがどのように示されるかを評価した．まず，次節で評価実験に用いるデータの説明をし，次に実際の評

表 4 実験に利用した Blog 記事（エントリ）数
Table 4 Number of blog articles used to experiment.

注目キーワード	タイプ	記事数
レストラン イタリアン	体験	122
	非体験	95
iPod nano	体験	104
	非体験	110
ヘルシオ	体験	93
	非体験	105

価実験について説明する．

5.1 正解データとパラメータ設定

まず，実験で利用する評価用データについて説明する．本実験では，先に紹介した goo リサーチ[2] の「ユーザが書き込む Blog の情報を参考にして，商品やサービスに興味を持ったり，購入・利用したりする」という結果から，Blog 記事（エントリ単位）を解析対象文書として用いる．Blog 記事は，goo のブログ Search[15] でキーワード検索を行った結果を用いる．検索キーワードは，体験した情報が重視されるような内容で，かつ Blog に記述されることが多いと想定される「食事・携帯端末・家電」の 3 タイプから「レストラン イタリアン」「iPod nano」[⋆1]「ヘルシオ」[⋆2]を選定した．「iPod nano」や「ヘルシオ」は，データ取得当時に販売されて間もない商品名であり，「レストラン」は，体験して分かるサービスや味に関係する語句である．これら 3 つのキーワードに関係した体験情報を消費者に示すことは，自ら体験しなくても他者の体験を知ることで物の価値が分かるようになることから，消費者にとってより効果的であると判断した．

検索結果の上位からキーワード（注目キーワード）に関わる体験情報が記述されている記事と体験情報が記述されていない記事，両者判断困難な記事に人手で分類し，体験情報が記述されている記事と体験情報が記述されていない記事の 2 種類（約 100 件ずつ）を評価用データとした．データの数量は，**表 4** のとおりである．体験情報か否かは，2 人の人手によって主観的に判断した．

体験情報と判断される記事は，たとえば例 (i)，(ii) のようなものが存在する．これらは，記事の書き手がイタリアンレストランへ行ったことや iPod nano を購入したという体験をしていると判断できるものとして選別されている．また，体験情報でないと判断される記事は，例 (iii)，(iv)，(v) のようなものが存在する．例 (iii) は，宣伝のような記事であり，体験情報ではない．例 (iv) は，iPod nano のコネクタについて Web 上のサイトから情報を収集し報告しているが，書き手は体験していないため，体験情報ではない．また例 (v) は，本人がパスタを作ってはいるが，イタリアンレストランについての体験ではない．

4 章で述べた I. 算出領域について次のように設定し，実験を行った．

> 例 (i) 初めて連れて行ってもらった***イタリアンレストラン***でメニューにない特別スパをオーダー．そしたら，すっごいスパゲティ出てきました．やみつきになりそぉ〜...

> 例 (ii) 今日はね，待ってましたの ***iPod nano*** ちゃんがうちに来たよ．黒にしようと思ったけど，PC に合わせて白にしたよ...

> 例 (iii) **ヘルシオ**は，おいしくできてしかもヘルシー！余分な油を落とします．水蒸気を加熱して 100℃ 以上の高温状態にした無色透明の気体を利用しています！...

> 例(iv) ***iPod nano*** は，iPod とコネクタが微妙に違うようです．詳しくは，この記事を http://.... ということで皆さんも注意してね！...

> 例 (v) 新しくできた***イタリアンレストラン***，に行ってみたいなぁ．でも，今日はおうちでパスタ．イタリアンレストランで作ったみたいにおいしかったよ...

I．算出領域：X，Y

注目キーワードに関係する体験表現は，注目キーワードよりも前であれば，「昨日**買った**ヘルシオは，〜」（体験表現を下線太字で示した）というように，注目キーワードに係る連体修飾として直前に出現することが想定される．また注目キーワードよりも後ろであれば，「ヘルシオでアップルパイを**焼きました**．とっても簡単で**使いやすかった** です.」のように，注目キーワードから少し離れた位置にも出現することが想定される．連体修飾であれば，注目キーワード直前の 10 文字程度，注目キーワードが後ろの文章に係る領域であれば 100 文字程度（2〜3 文程度）以内と想定し，算出領域を次のように設定し，それぞれの精度を確認する．

X：0，10，20 バイトの 3 種類

Y：50，100，150，200 バイトの 4 種類

形態素数や文，段落などで範囲を指定し，解析を行う例もあるが，本実験では，解析対象文書が Blog で

⋆1 iPod nano は，米国アップルコンピュータ社の登録商標である．
⋆2 ヘルシオは，シャープ（株）の登録商標である．

あることから，形態素，文の切れ目，段落の区切りの判定が困難である場合があることに鑑み，バイトでの指定とした．

5.2 評価実験

前節で示した正解データで，スコア算出式 (3) を利用し評価実験を行う．解析対象となる Blog 記事を形態素解析し，4 章で述べた体験表現ルールに適合する語句があるかどうかを事前に検査しておく．このデータを基に，スコア算出式 (3) を利用し，注目キーワードごとの体験らしさのランキングを行う．

スコア算出式 (3) の各パラメータは，それぞれの算出領域の組合せ（$X = 0$ $Y = 50/100/150/200$, $X = 10$ $Y = 50/100/150/200$, $X = 20$ $Y = 50/100/150/200$）において，4 章で示した「II. 注目キーワードと体験表現の共起出現頻度：$II(T_i)$」と「III. 注目キーワードと体験表現の距離：$III(T_{ij})$」のすべての組合せ（II のみ，III のみ，II と III 両者）を用いる．1 つの注目キーワードに対して 36 通りの組合せが存在する．II のみの場合は，$III(T_{ij})$ の各体験表現の種類 T_i の総和を 1（注目キーワードと体験表現との距離を考慮しない）とし，III のみの場合は，$II(T_i) = 1$ として算出する．

これらの結果，体験情報抽出システムが人手でラベル付けした体験情報を抽出できた割合（適合率）を算出する．

6. 評価結果

我々は，体験らしさに基づく解析対象文書のランキングを狙いとするため，体験情報表示結果の網羅性ではなく，ランキング上位に体験情報が高い割合で提示することが要求される．そこで本実験では，ランキング上位の適合率によって，提案手法の精度を評価する．先に述べた 3 つの注目キーワードについての実験結果をグラフを用いて説明する．適合率は式 (4) を利用し，提案する体験情報抽出システムが自動算出の結果提示した記事の中で，人手で体験情報であるとラベル付けした記事が含まれる割合を求めた．

$$\text{適合率} = \frac{\text{人手でラベル付けした体験情報の中でシステムで自動取得できた記事数}}{\text{システムで自動取得した全記事数}} \quad (4)$$

6.1 算出領域別の適合率の変化

検索結果が出力された際に，システムユーザが見る結果の範囲を上位 10 件程度と想定し，上位 10 位までの適合率を検査した．各算出領域での上位 10 件までの適合率の変化は**図 2** のグラフのようになった．図 2 は，3 つの注目キーワードの平均値である．縦軸に適合率，横軸は，各算出タイプの組合せを示す．棒グラフの凡例は，注目キーワードよりも前の算出領域 X を指し，0/10/20 バイトである．また，横軸の「Num_f, Num_d, Num_df」は，Num が注目キーワードよりも後ろの算出領域 Y を指し，50/100/150/200 バイトの結果である．また，Y の後ろに並ぶ「f, d, df」は，スコア算出式 (3) において，

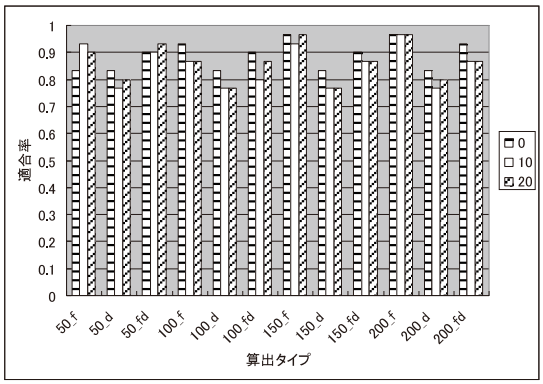

図 2 各算出タイプにおける適合率（平均）
Fig. 2 Precision in each calculation type (average of key words).

- f：式 (2) の各体験表現の種類 T_i の総和を 1（注目キーワードと体験表現との距離を考慮しない）とし，II. 共起出現頻度の式 (1) のみを算出

$$if(X \le T_{ij} \le Y) : S(T) = \sum_{i=1}^{m}(II(T_i)) = \sum_{i=1}^{m}\left(\frac{T_i\_df}{N}\right) \quad (5)$$

- d：式 (1) $II(T_i) = 1$ とし，III の距離の式 (2) のみを算出

$$if(X \le T_{ij} \le Y) : S(T) = \sum_{i=1}^{m}\left(\sum_{j=1}^{k} III(T_{ij})\right) = \sum_{i=1}^{m}\left(\sum_{j=1}^{k}\frac{1}{T_{ij}\_dis}\right) \quad (6)$$

- df：II，III の両者を算出（式 (3) そのまま）

の 3 タイプを示している．

この結果では，どの算出タイプにおいても適合率が 75%を超えた．特に $X = 0$ では，80%以上の適合率を示した．そして「50_f」「50_fd」を除いた算出領域 $X = 0$ では，つねに最も高い適合率を示した．その中でも，$Y = 50$ を除いた算出領域では，算出タイプ f（II の式のみを算出：式 (5)）が他と比べて高い適合率

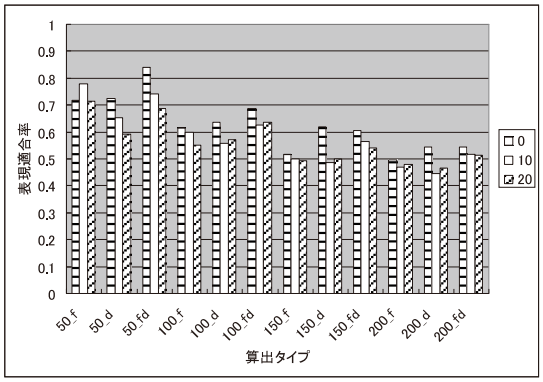

図 3 各算出タイプにおける表現適合率（平均）
Fig. 3 Expression Precision in each calculation type (average of key words).

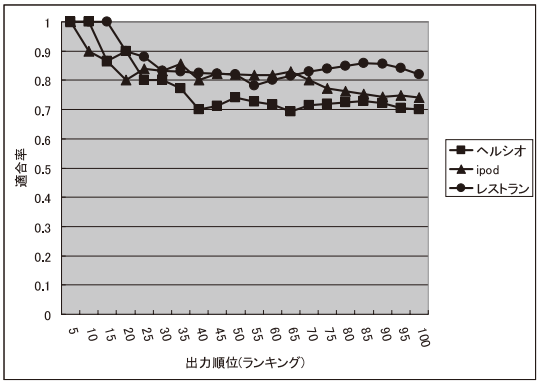

図 4 注目キーワードごとの適合率の推移
Fig. 4 Precision of each key word.

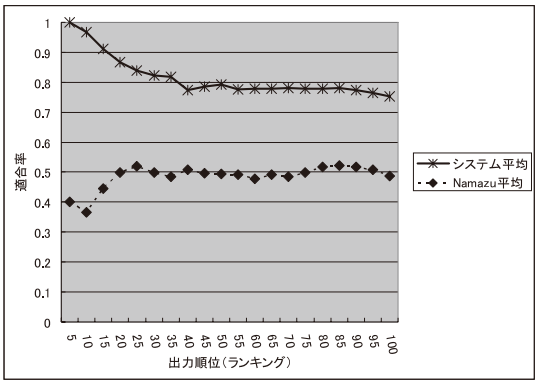

図 5 キーワード検索との適合率の比較（平均）
Fig. 5 Precision of general key word search and proposal system.

を示した．

さらに上位 10 件において，体験表現の抽出精度を確認するために，次のことを行った．各算出領域内で注目キーワードに関係する体験表現を人手でラベル付けし，システムで自動取得した場合の表現適合率を確認した．ここでの人手の作業は 2 人で行い，両者合意のとれた表現を体験表現と判断した．表現適合率は，式 (7) を利用した．

$$\text{表現適合率} = \frac{\text{人手でラベル付けした体験情報の中でシステムで自動取得できた表現数}}{\text{システムで自動取得した体験表現数}} \quad (7)$$

図 3 に示すとおり，「50_f」を除いたすべてにおいて $X = 0$ が最も高い表現適合率を示しており，注目キーワードよりも前に出現する体験表現は，5.1 節での想定よりも精度向上に寄与していないことも分かった．これは，本システムが 1 文ごとに解析しておらず，算出領域というブロックで評価していることと形態素解析がうまく行えていないことが原因の 1 つであると考えられる．たとえば「今日は，秋葉原へ**行った**よ でも iPod nano 見に行けず，さみし〜」というような文章があったとき，文の切れ目がうまく抽出できた状態で，1 文単位での評価を行えていれば，iPod nano より前に出現する，「行った」は体験表現ルールにあてはまっても，iPod nano に関する体験を表す表現ではないと判断でき，誤認識することはない．

6.2 注目キーワードごとの適合率の推移

解析領域が少ない方がシステムに与える負荷が少ないことから，最も適合率が高い算出タイプの中で，算出領域が最も狭い「$X = 0$，$Y = 150$，f，(150_f)」（図 3 の結果より）において，各注目キーワードの適合率の変化を上位 100 件までについて確認した．

図 4 に注目キーワードごとの各ランキングにおける適合率の推移を示す．上位 30 件までは，どの注目キーワードも 80%の精度を保っている．そして，上位 100 位までを見ても 70%の精度が出ている．

本手法は構文解析などを用いず，体験表現ルールに基づき抽出した体験表現と注目キーワードとの関係を「共起」と「バイト数指定の算出領域」を用いてシンプルに算出しているが，上位 5 件における適合率は 100%となった．

6.3 キーワード検索と体験情報抽出システムの比較

本提案システムとキーワード検索でよく用いられる Namazu 検索[16] との比較を行った．図 4 で示した本提案システムの結果の平均値と，同じ 3 つの注目キーワードで Namazu 検索を行った結果の平均値を**図 5** に示す．Namazu 検索では，上位 20 件くらいまでは不安定ではあるものの，およそ 50%の確率で体験情報と非体験情報が混在して出力されている．つまり，Namazu のような体験情報を考慮しないキーワード検索では，情報が混在し体験情報のみを探すことが難し

いことを示している．本提案システムと比較すると，上位 10 件では最も大きい 60%程度の差が現れ，上位 100 件まで見ても 25%の差があることが分かった．このように本提案システムを利用することで，高割合で体験情報が上位にランキングされることが分かる．

6.4 実験の考察

本提案手法では，注目キーワード周辺の体験表現の有無を確認し，体験情報であるか否かを判定するため，まずは，システムで自動取得した体験表現が正しく取得できているか否かについて考察する．

例 (vi)〜(viii) に本実験でランキング上位に出力された Blog 記事と類似した文章例を示す．例文では，注目キーワードを***斜体太字***で，システムが判定した体験表現を**下線太字**で示した．例 (vi) でシステムが判定した体験表現は，すべて注目キーワードに関係する体験表現である．例 (vii) では，「**買った**」は iPod nano に関係する体験表現である．そして，「壊れ**かけた**」や「**使っていた**」，「刺さ**れた**」は MD プレーヤに関係する体験表現であり，「**決心した**」は買い換える行為自体に関係する語句であるため，iPod nano に関係する体験表現とはいえない．しかしながら，iPod nano を買うに至った経緯が書かれており，まったく関係のない内容ともいいがたい．例 (viii) では，ヘルシオがレンジにいい代わっているが，「**欲しかった**」や「**買ってしまった**」は，ヘルシオに関係する体験表現である．

> 例 (vi) ***イタリアンレストラン***に**行った**のだ．先に電話で**予約してた**ので待たないで**入れた**のだ〜パスタとピザのランチ**食べた**のだ．ボリュームあったのでお腹一杯に**なった**のだ...

> 例 (vii) この間 ***iPod nano*** を**買った**．壊れ**かけた** MD プレーヤを四年間**使っていた**が，友達に机から落とされ，とどめを刺さ**れた**ので買い替えようと**決心した**...

> 例 (viii) **ヘルシオ**を買うことに・・・どうしてもこのレンジは**欲しかった**ので高めの値段ですが思わず**買ってしまった**...

本システムでは，構文解析を行っていないが，このようにして，ランキング上位に出力される Blog 記事は，例 (viii) のように直接注目キーワードに係っていなくても体験表現として正しく取得し，体験情報と判断できていることが分かる．

次にスコア算出式について考察する．本実験では，「II. 注目キーワードと体験表現の共起出現頻度」と「III. 注目キーワードと体験表現の距離」を考慮したスコア算出を行った結果，II を利用して算出した式 (5) の方が適合率が高くなった．算出領域内において，他の解析対象文書と共起数の多い体験表現は，注目キーワードに関係した体験を表す語句である可能性が高いことが改めて確認できた．距離よりも共起が効いているのは，他文書の書き手も注目キーワードに対して同様の体験をしていることが多いことを表している．しかしその反面，この手法では注目キーワードに対するごく少数の人の体験は，ランキング下位になる．

ただし，Blog のような移り変わりの早い文書において，「II. 注目キーワードと体験表現の共起出現頻度」を利用することにより，解析した文書群における時間的な流行を知ることができる．たとえば，ヘルシオの場合は，買いたいと思っていたが値段を気にしている人が多く，安くなったので買ったという理由が付随していることが多い．また，「買った」という体験表現の次に共起出現頻度が高かったのが「焼いた」などのヘルシオを使って料理した記事で，まだまだヘルシオが目新しく，それで料理したことをいち早く伝えたいというような雰囲気が現れている．また，ある商品の発売前，発売後，マイナーバージョンアップ，次機種発売，...などのようなイベントごとに見ることで，消費者の行動の変化も観察できる．このように移り変わりの早い文書を解析すると，注目キーワードに対する時期ごとの流行のスタイルが発見できる可能性がある．

7. ま と め

本論文では，体験情報を抽出することは，今後のユーザの消費活動やマーケティングに影響を与えると考え，体験情報抽出システムを提案し，評価実験を行った．

本手法は，構文解析などを用いず，体験表現ルールに基づき抽出した体験表現と注目キーワードとの関係を共起や距離，算出領域を用いてシンプルに体験らしさを測った．実験では，調査した注目キーワードの数や解析文書数は多いとはいえないが，その中では出力結果上位 5 件において 100%，100 件において 70%以上の適合率で体験情報を提示可能であった．また，体験情報を考慮しないキーワード検索と比較すると，上位 10 件では 60%，上位 100 件では 25%，本提案システムの適合率が上回り，高割合で体験情報を提示可能なことも確認した．さらに，体験情報をランキング上位に上げるためには，ある一定の算出範囲内において，注目キーワードと体験表現の共起関係を調べることが効果的であることも分かった．また，ランキング

上位に出力される Blog 記事は，直接注目キーワードとの係り受け関係がなくても，システム側で体験表現として正しく取得し，体験情報と判断できることが分かった．

今後，ランキング精度をさらに向上させるためには，体験表現自体の抽出精度の向上も必要である．Blog のような移り変わりの早い文書では，注目キーワードに対する時期ごとの流行スタイルが発見できる可能性があることから，本論文で設定した体験表現ルールの精度確認に加え，確認するキーワードを増加させ，利用分野も考慮した体験表現の抽出精度についても検討していく予定である．

参 考 文 献

1) 価格.com ニュースレター 2005 年 10 月 31 日. http://kakaku.com/info/press_release/20051031.pdf
2) goo リサーチ：ネット上の口コミ情報と広告に関する調査. http://japan.internet.com/research/20061128/1.html
3) 立石健二，石黒義英，福島俊一：インターネットからの評判情報検索，情報処理学会研究報告，Vol.2001, No.69, pp.75–82 (2006).
4) 廣嶋伸章，山田節夫，古瀬　蔵，片岡良治：評判検索におけるクエリ依存型の評価極性付与，情報処理学会研究会報告，Vol.2006, No.124, pp.129–134 (2006).
5) 佐藤吉秀，関口裕一郎，川島晴美，奥田英範：投稿記事間の “ばらつき” を利用したブログ分類手法，情報処理学会第 68 回全国大会，5C-2, pp.2_99–2_100 (2006).
6) goo リサーチ：1 万人のインターネットショッピング意向調査. http://research.goo.ne.jp/Result/0009op29/01.html
7) 乾　孝司，奥村　学：テキストを対象とした評価情報の分析に関する研究動向，自然言語処理，Vol.13, No.3, pp.201–241 (2006).
8) 藤村　滋，豊田正史，喜連川優：Web からの評判および評価表現抽出に関する一考察，信学技報，Vol.104, No.177 (2004).
9) Turney, P.D.: Thumbs Up or Thumbs Down? Semantic Applied to Unsupervised Classification of Reviews, *Proc. 40th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp.417–424 (2002).
10) Liu, B., Hu, M. and Cheng, J.: Opinion Observer: Analyzing and Comparing Opinions on the Web, *The 14th International World Wide Web Conference* (2005).
11) 小林のぞみ，乾健太郎，松本裕治，立石健二，福島俊一：意見抽出のための評価表現の収集，自然言語処理，Vol.12, No.3 (2005).
12) 廣嶋伸章，山田節夫，奥　雅博：概念ベースを用いた Web ページからの評価項目の自動抽出，言語処理学会第 11 回年次大会発表論文集，pp.428–431 (2005).
13) Kurashima, T., Tezuka, T. and Tanaka, K.: Mining and Visualization of Visitor Experiences from Urban Blogs, *Proc. 17th International Conference on Database and Expert Systems Applications* (*DEXA 2006*), Krakow, Poland (2006).
14) 池田佳代，定方　徹，奥　雅博：体験表現を手がかりにした Blog の評判情報判定方法の検討，電子情報通信学会第二種研究会資料，WI2-2005-36, pp.47–52 (2005).
15) goo ブログ Search. http://blog.goo.ne.jp/
16) 全文検索エンジンシステム Namazu. http://www.namazu.org/

(平成 19 年 5 月 14 日受付)
(平成 19 年 11 月 6 日採録)

池田　佳代（正会員）

1996 年群馬大学工学部情報工学科卒業．同年日本電信電話（株）入社．以来，知的教育システムの教材オーサリングシステムの研究開発を経て，NTT サイバーソリューション研究所において，CGM における情報抽出の研究開発に従事．現在は，NTT コミュニケーションズ（株）にて情報推薦に関する研究開発に従事．1998 年本学会第 56 回全国大会奨励賞受賞．

田邊　勝義（正会員）

1985 年横浜国立大学工学部情報工学科卒業．1987 年同大学大学院修士課程修了．同年日本電信電話（株）入社．以来，パターン認識，類似画像検索，車番号認識，眼底画像合成の研究に従事．現在，NTT サイバーソリューション研究所主任研究員．電子情報通信学会，映像情報メディア学会各会員．

奥田 英範

1988 年東京大学大学院工学系研究科修士課程修了．同年日本電信電話（株）入社．NTT サイバーソリューション研究所主幹研究員．1994 年スタンフォード大学コンピュータ科学科修士課程修了．現在，CGM における情報抽出の研究開発に従事．電子情報通信学会，映像情報メディア学会各会員．

奥 雅博（正会員）

1984 年大阪府立大学大学院工学研究科電子工学専攻博士前期課程修了．同年日本電信電話公社（現 NTT）に入社．自然言語処理技術，特に日本語処理技術の研究開発に従事．現在 NTT サイバーソリューション研究所において，ブロードバンドインターネットサービスに関する研究開発に従事．博士（工学）．電子情報通信学会，言語処理学会各会員．