体験表現を手がかりにしたBlogの体験情報の抽出

池田　佳代　　田邊　勝義　　奥田　英範

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　〒239-0847 神奈川県横須賀市光の丘1-1

E-mail:　{ikeda.kayo, tanabe.katsuyoshi, okuda.hidenori}@lab.ntt.co.jp

あらまし　Blogに代表されるような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。そこで本稿では、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、評価実験を行なった。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に関係するBlogを解析した結果、通常のキーワード検索と比較し、提案システムの方が約25%多く体験情報を抽出できることを確認した。

キーワード　体験表現，体験情報抽出，Blog

Using Experience Expressions to Extract Experience Information from Blogs

Kayo IKEDA　　Katsuyoshi TANABE　and　Hidenori OKUDA

NTT Cyber Solutions Laboratories, Nippon Telegraph and Telephone Corporation

1-1 Hikarinooka, Yokosuka-Shi, Kanagawa, 239-0847 Japan

E-mail:　{ikeda.kayo, tanabe.katsuyoshi, okuda.hidenori }@lab.ntt.co.jp

Abstract　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.

Keyword　Experience Expressions, Experience Identification, Evaluation Expressions, Blog

1. はじめに

近年、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、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等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、その利用者数も、日を追うごとに増加している。このようなBlog記事から意見の抽出や話題語による内容の分類等の試みが行われるようになってきた[3,4]。

評判情報が書かれる物事（財）は、経験財と探検財に分けることができる[5]。探検財とは、消費者が購入前に調べる事で、そのもの自体の良し悪しが分かる財を指し、経験財とは、使ってみて初めて良し悪しがわかるような財を指す。gooリサーチにおいても、「経験財・探検財の両者で、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体験情報は、評判情報と組合せたときに、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出来るため、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や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までには、いたっていない。一方、所望のキーワードに関わる体験情報は、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鑑みて検索することは出来ず、単なるキーワードマッチングによる検索になってしまうことから、探すことが困難である。

本研究では、分野に依存せず、ユーザの体験による情報を抽出する試みの一つとして、体験から得た情報を記述する際の特徴を体験表現とし、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システム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、本稿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、肯定・否定文を “excellent”や“poor”との共起関係を統計量を用いて判断する研究 [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、その製品の評価軸を自動取得する手法 [10]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、文書から評価表現を抽出する精度を向上させようとすると、評価したい対象によって、評価表現を変化させる必要が出てくる。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[11]。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事は、コストが高くなる事から、半自動的に作成する研究が行われている[12]。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[13]。

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事ができる。

しかしながら、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。

体験情報に関する研究としては、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、観光情報を取得する試みがある[14]。Kurashima らは、地名に特化しているが、本研究では、幅広い分野に対応するため、ドメインに依存せずに体験を抽出することを考えている。

著者らは、体験表現と評価表現を組合せることで体験的評判情報を判定する試みを行った[15]。評価表現に加え体験表現を用いる事で、宣伝のような文書を除去する効果もあると考えられる。また、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出来るため、読み手の理解が深まると考えられる。しかしながら、著者らの研究[15]では、ユーザが注目するキーワードに対する体験情報を抽出することは、考慮していなかった。

評判情報については、既に広く研究されているため、本研究では体験に特化し、ユーザが注目するキーワードに対する体験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、本研究で用いた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

体験情報を抽出する一方法として、本研究では体験を記述する際に現れる「体験表現」に着目する。体験表現とは、主に自発的な動作を表す動詞の過去形・進行形・動作を表す名詞が該当する。書き手の体験の結果、得られたであろう感想を表すような形容詞の過去形等も含まれる。自己の体験ではなく、他の状態を表すような「死ぬ、消える、終わる」などは含まない。表 1 に体験表現の 5 つのタイプとその例を示す。

表 1　体験表現のタイプ

	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、挑戦した、留学した、体験した、…
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、…
④	動詞（名詞の動作を表す単語含む）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、…

3.2. 体験情報について

前節で述べたような“体験を記述する際に現れる体験表現”が含まれてい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。

表 2　本稿で扱う体験情報の例

	例文	判定
a)	毎年、○×温泉へ***行っています***。	○
b)	一月に一度は、△レストランで***食事をしています***。	○
c)	最近、渋谷に新しく出来た△レストランへ***行ってきました***。	○
d)	昨日の温泉は、とても***良かった***です。	○
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。	×

例文 a),b),c)は、表 1 の③書き手が行動した事もしくは、④行為を継続中である事について、「行っています」「食事をしています」「行ってきました」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。また、例文 d),e)も、表 1 の⑤書き手の経験から得た感想を「良かった」「悪かった」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ことに加え、f)は、マシンが安いという条件を示しているだけで、書き手が体験したものかどうかは、この 1 文だけでは特定できないため体験情報ではない。また g)についても、アメニティと料金という条件を示しているだけで、書き手が体験したものかどうかは特定できないため体験情報ではない。体験情報は、複数文にわたるものもあれば、表 2 の例のように、1 文で表現されるものもある。

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に、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、どのような体験表現があるかを検査する。ランタイム処理では、システム利用者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。

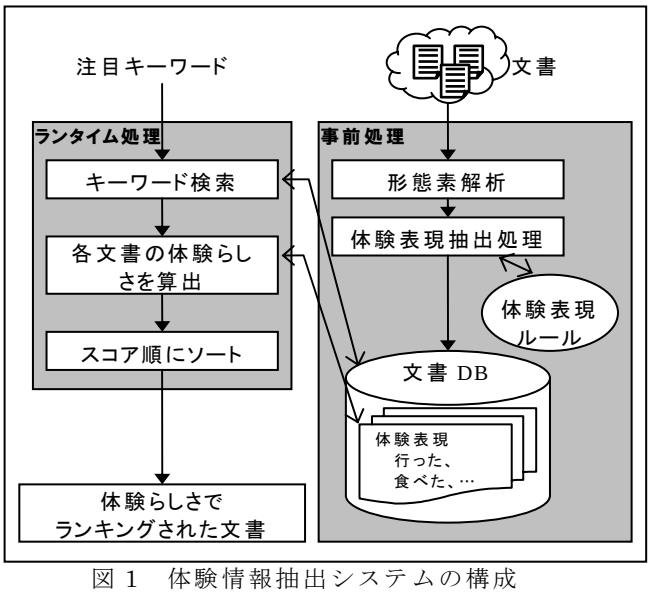


図 1　体験情報抽出システムの構成

4.2. 体験表現の検査

まず、体験情報抽出システムは、解析対象文書の中にどのような体験表現があるかを検査する。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①②に示したような「～してみた、～したことがある」というような言い回しや表 1 の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、20 種類の品詞の組み合わせで設定し、表 3 に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。

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品詞の並びが存在した場合は、体験表現があるとみなす。解析対象文書すべてから体験表現の有無を検査する。

本ルールは、先に紹介した著者らの研究[15]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。

表 3　体験表現のルール例

	表現タイプの説明	ルールの例
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」}　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

4.3. 注目キーワードと体験情報の関係

システム利用者によって、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言えるスコアである。スコアの算出式は、次のようなパラメータを用いる。

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の高い体験表現は、注目キーワードとの関係性が強いことが考えられる。例えば、「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が高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全てにおいて、注目キーワードと体験表現の共起が高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。式は、ある種類の体験表現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

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

を利用することとする。

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、本研究では、解析対象文書が Blog であることから、構文解析を行うコスト・誤解析のリスクを鑑み、単純な距離での判定とした。

本研究では、T_{ij}_disを注目キーワードとある体験表現T_{ij}との距離＝バイト数とし、その逆数の式(2)を利用することとする。

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

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。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数を指す。

$$if\,(X \le T_{ij} \le Y): S(T)=\sum_{i=1}^{m}\left(\frac{T_i\_df}{N}*\sum_{j=1}^{k}\frac{1}{T_{ij}\_dis}\right) \quad (3)$$

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の高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出来るか、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]「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[16]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、なおかつ 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」「ヘルシオ」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ため、1 章で紹介した経験財（体験して初めて価値がわかる財）に相当する。

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。

表 4　実験に利用した Blog 記事（エントリ）数

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

体験情報と判断される記事は、例えば例 1),2)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だろう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 3),4),5)のようなものが存在する。例 3)は、宣伝のような記事であり、体験情報ではない。例 4)は、iPod nano の問題について 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 5)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。

例 1）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 2）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 3）***ヘルシオ***は、おいしく出来てしかもヘルシー！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 4）***iPod nano***は、コネクタが廃止になっているとかで互換問題が発生しているようです。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 5）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。...

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。

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、～」というように、注目キーワードの直近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。よって、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こととする。

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を鑑み、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、評価実験を行った。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）をそれぞれの算出領域の組み合わせ（*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})」の全ての組み合わせ（Ⅱのみ、Ⅲのみ、ⅡとⅢ両者）で算出した。1つの注目キーワードに対して36通りの組み合わせが存在する。Ⅱのみの場合は、Ⅲ(T_{ij})=1とし、Ⅲのみの場合は、Ⅱ(T_i)=1として算出する。

4 章で述べた体験表現ルールを用いて、解析対象となる Blog 記事を形態素解析し、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。

6. 評価結果

本研究では、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ことから、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されることが要求されている。そこで本実験では、ランキング上位の適合率によって、システムの精度を評価する。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。

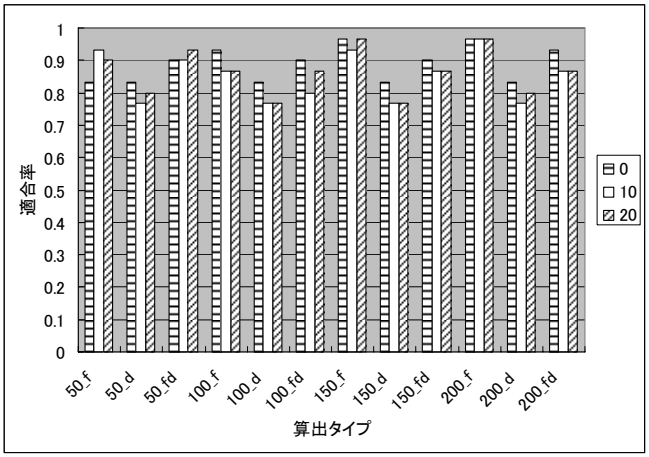
$$\text{適合率} = \frac{\text{人手でラベル付けした体験情報の中でシステムで自動取得できた記事数}}{\text{システムで自動取得した全記事数}} \quad (4)$$

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、システム利用者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グラフ1のようになった。グラフ1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、スコアを算出する際の式で *f*:Ⅱ（出現頻度）有効、*d*:Ⅲ（距離）有効、*df*:ⅡⅢの両者有効の 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 7.5 割を超えた。そして、「*50*_f」を除いた全てにおいて、算出領域が *X*=0 で、算出タイプ *f*（Ⅱ有効）が最も高い適合率を示し、8 割以上の精度が出た。

さらにグラフ 1 からは、注目キーワードよりも前に出現する体験表現は、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、本システムが一文毎に解析しておらず、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。例えば「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当てはまっても、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。



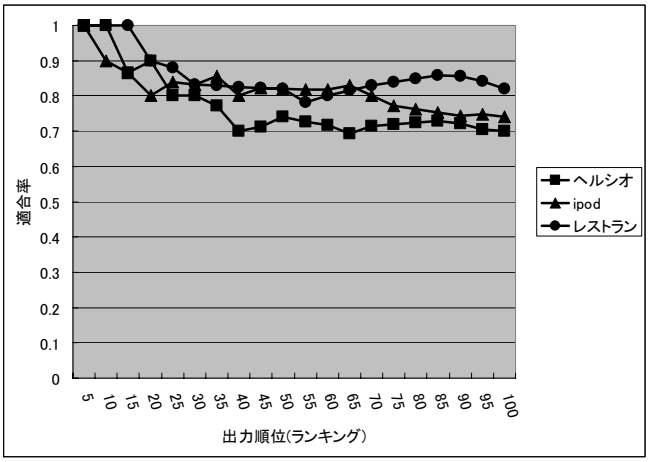
グラフ 1　各算出タイプにおける適合率（平均）

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くないことから、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*)」（グラフ 1 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。

グラフ 2 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 8 割の精度を保っている。そして、上位 100 位までを見ても 7 割の精度が出ていることが分かる。また、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

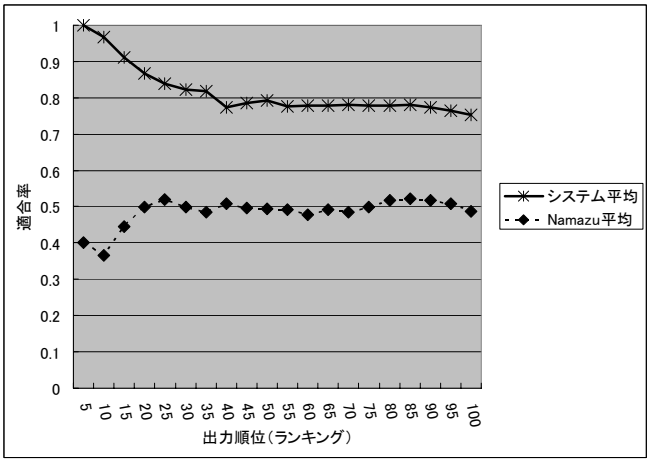
の関係を「共起」と「バイト数指定の算出領域」を用いてシンプルに算出しているが、上位 5 件における適合率は 100%となった。



グラフ 2　注目キーワードごとの適合率の推移

6.3. 検索結果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[17]との比較を行った。グラフ 2 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グラフ 3 に示す。



グラフ 3　キーワード検索との適合率の比較

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ることが分かる。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ということが確認できる。これに比べ、本体験情報抽出システムは、上位 40 件までは、適合率が下降しても 8 割程度を保ち続け、それ以降は緩やかな下降をし、7.5 割の精度で落ち着く。このように本体験情報抽出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

まず、本実験で上位にランキングされた Blog 記事の文章について考察を行う。

例 1)～3)に実際に出力された 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 1)でシステムが判定した体験表現は、全て注目キーワードに関係する体験表現である。例 2)では、「**買った**」は iPod nano に関係する体験表現である。「壊れ**かけた**」や「**使っていた**」、「刺さ**れた**」は MD プレーヤー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言えない。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言いがたい。例 3)では、ヘルシオがレンジに言い代わっているが、「**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。

本システムでは、構文解析を行っていないが、例 3)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取得し、体験情報と判断できていることが分かる。

例 1）***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 2）この間***iPod nano***を**買った**。壊れ**かけた** MD プレーヤーを四年間**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 3）***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**...

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が、結果的にⅡの式(2)を利用して算出した場合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が高い体験表現は、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という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ということだろう。しかしその反面、この手法では注目キーワードに対する極少数の体験は、ランキング下位になってしまう。

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出来る。たとえば、ヘルシオの場合は、買いたいと思っていたけれど、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、...などに対する消費者の行動の変化も観察できるだろう。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毎の流行のスタイルが見つかることが考えられる。

7. まとめ

本稿では、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言えないが、その中では出力結果上位 5 件において 100%、100 件において 75%の適合率で体験情報を提示することが出来ることを確認した。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 100 件において提案システムの適合率が約 25%上回り、高い割合で体験情報を提示することが出来ることも確認した。さらに、体験情報をランキングの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を調べることが効果的であることも分かった。

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。本稿で設定した体験表現ルールの精度確認に加え、体験表現は、利用分野に左右されることが考えられるため、確認するキーワードを増加し、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。

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] 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 自然言語処理研究会, 2006-NL-176, pp129-134, (2006.11)

[4] 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, (2006.3)

[5] Philip Nelson,“Information and Consumer Behavior”, Journal of Political Economy, 78(2),311-329, (1970)

[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” Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp417-424, (2002)

[10] B. Liu, M. Hu, J. Cheng, “Opinion Observer: Analyzing and Comparing Opinions on the Web”, The 14th International World Wide Web Conference, (2005)

[11] 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”, 情処学会研究報告, NL-144-11, pp75-82, (2001)

[12] 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)

[13] 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, pp428-431, (2005)

[14] Takeshi Kurashima, Taro Tezuka, and Katsumi Tanaka, “Mining and Visualization of Visitor Experiences from Urban Blogs”, Proceedings of the 17th International Conference on Database and Expert Systems Applications (DEXA 2006), Krakow, Poland, (2006.9)

[15] 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, (2005.9)

[16] goo ブログ Search, http://blog.goo.ne.jp/

[17] 全文検索エンジンシステム Namazu, http://www.namazu.org/