

大規模テキストからの経験マイニング

倉島 健[†] 藤村 考[†] 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: †{kurashima.takeshi,fujimura.ko,okuda.hidenori}@lab.ntt.co.jp

あらまし 本研究の目的は、非構造データであるブログに記述された人間の経験を構造化/知識化することである。経験とは、状況（時間、空間）、行動（動作、対象）、主観（評価、感情）とから成る情報であり、本稿では評価を除く 5 要素（時間、空間、動作、対象、感情）を抽出する手法を提案する。さらに、感情を 8 カテゴリーに分類することで、ある経験が動作主にとって成功だったか失敗だったかを導き出す。また、得られた大量の経験情報集合から、相関ルール抽出技術を用いて、状況、行動、主観との間の関係をルール形式で抽出する。提案手法を実装したシステムにより、実際に 2007 年 1 月から 5 月までの約 4800 万件のブログ記事から経験情報を抽出した。さらに、相関ルールの「興味深さ」を評価する様々な評価尺度を用いて、抽出データを多角的に分析し、その中からいくつかの興味深いルールを発見するに至った。

キーワード テキストマイニング, 知識発見, Blog

Mining Experiences from Large-scale Blog Entries

Takeshi KURASHIMA[†], Ko FUJIMURA[†], and Hidenori OKUDA[†]

[†] NTT Cyber Solutions Laboratories, NTT Corporation

1-1 Hikari-no-oka, Yokosuka-shi, Kanagawa, 239-0847 Japan

E-mail: †{kurashima.takeshi,fujimura.ko,okuda.hidenori}@lab.ntt.co.jp

Abstract An important characteristic of Weblogs(blogs) is that they contain many descriptions of people's experiences in the real world. This paper proposes a method for extracting people's experiences from large-scale blog entries and also a method for mining association rules between location, time, activity, and emotion. An activity consists of action and its object. We also categorize people's emotions into nine types, and classify each experience into success and failure based on the emotion categories. We constructed a system that mines association rules from about 48 million blog entries, and analyzes data from many directions using several interesting measures for data mining. As a result, we successfully found interesting rules about people's activities.

Key words Text Mining, Knowledge Discovery, Weblog

1. はじめに

近年、ブログや SNS、掲示板などのいわゆる CGM と呼ばれるメディアの普及が著しい。CGM の重要な特徴は、人々の主観的な記述が多いことである。これらの情報は、広告主や企業が発信する情報とは異なる消費者の「生の声」として価値があり、マーケティングや、企業経営、消費行動等の分野で活用したいという要望は高い。

一方で、CGM をはじめとした Web 上に存在する大多数の情報は、自然文で記述された非構造データである。また、ブログや RSS で配信される情報は膨大であるが、玉石混交であるため、奥村ら [1] の話題語抽出システム blog watcher や、kizasi.jp [2]

等のサービスに代表されるように、「情報が新鮮である」という観点での利用がほとんどである。古くなった情報はやがて破棄されてしまう運命にある。

このような現状を踏まえ、我々は、ブログという非構造データを対象とし、従来捨てられてしまっていた、いわば個人の「知」そのものを再利用可能な形に構造化することを目指す。構造化データとは、コンピュータが処理できるように、その属性や意味を規定した情報である。本研究においては、ブログに記述された情報の中でも特に人間の経験情報の価値に着目した。経験情報を、状況（時間、空間）、行動（動作、対象）、主観（評価、感情）とから成る情報と定義する。ブログから経験情報を抽出（構造化）し、さらには大量の経験情報の中から、状況と行動と

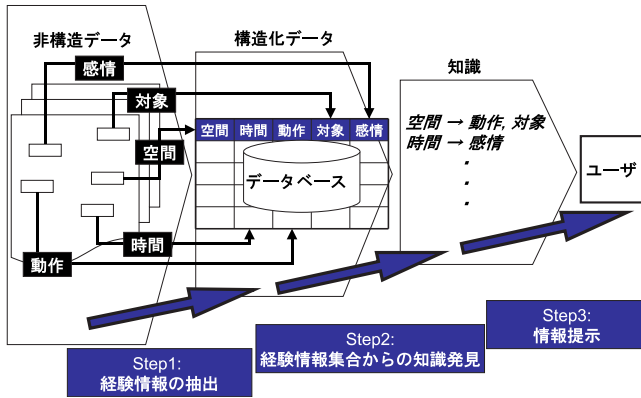


図1 非構造化データを構造化する処理の全体像

主観との間の興味深い関係をルール形式で抽出（知識化）することができれば、人間が、他人の経験を生かし、自らの経験なくしてより適切な行動を選択可能になる。本稿においては、経験情報の中でも特に5要素（時間、空間、動作、対象、感情）を抽出する手法の概要を述べる。さらに、人間の感情を8カテゴリに分類することで、それぞれの経験情報が、動作主にとって成功だったか、失敗だったかを導き出した。また、得られた大量の経験情報集合から、相関ルール抽出技術を用いて状況と行動と主観との関係をルール形式で抽出する。図1に、提案手法の全体像を示す。提案手法を実装し、実際に2007年1月から5月までの約4800万件のブログ記事から経験情報を抽出した。さらに、相関ルールの「興味深さ」を評価するリフト、 χ^2 値[3]や、J-measure[4]などの様々な評価尺度を用いて、抽出データを多角的に分析し、その中からいくつかの興味深いルールを発見することに成功した。

以下、本稿の構成を記す。2章では、経験情報の定義について、3章では、関連研究について述べる。4章では、ブログテキストからの経験情報抽出手法について述べる。5章では、経験情報集合からの知識発見手法を、6章では、作成したシステムの構成について、7章では、評価実験の結果について述べる。8章では、まとめと今後の課題を述べる。

2. 経験情報の定義

本研究においては、人間の経験は以下の3要素から構成されるとする。

- (1) 状況：行動を行った時間と空間。
- (2) 行動：人間が行う動作とその対象。
- (3) 主観：動作を伴う対象に対する評価と、一連の行動を行った結果、動作主が抱いた感情。

それぞれの要素はさらに細分化でき、実際には以下に示す情報を人間の経験を表現する最小要素としてテキストから抽出する。

$E = \{ \text{時間, 空間, 動作, 対象, 評価, 感情} \}$

例えば、「昨日、嵐山で紅葉を見ましたが、とてもきれいで感動しました…」という文章は{時間, 空間, 動作, 対象, 評価, 感情} = {昨日, 嵐山, 見る, 紅葉, きれい, 感動}のように、経験情報で表現できる。なお、「誰が」という「動作主」に関する情報は、通常、テキスト中に明示されないため、今回の定義には含めない。本稿では、評価を除く、5要素について考える。

3. 関連研究

評判情報抽出の分野においては対象、属性、評価という3つ組を抽出することが主要な技術課題となっている。立石ら[5]は対象、属性、評価に関する共起パターンを介して、属性表現と評価表現をブーツトラップ的に抽出する手法を提案した。Bingら[6]も、3つ組のうちの属性要素を自動生成する手法を提案している。従来の評判情報抽出が、人名や組織名、商品名等の「対象」を軸とした情報抽出なのに対し、我々は、「いつ（時間）」、「どこで（空間）」、「何をする（動作、対象）」という人間の「行動」を軸とした経験そのものの情報抽出を行う。

感情要素の抽出において、福原ら[7]は、新聞記事から人手で感情語を収集し、感情辞書を作成している。また、熊本ら[8]は、2つの尺度「悲しい-嬉しい」、「怒る-喜ぶ」に対する評価値(0~1)を含む感情表現の抽出手法を提案した。本研究においては、これらの知見を最大限生かして、感情を構成する8カテゴリを設定した。また、安田ら[10]は、一般的にテキスト上に直接的に記述されることはないブログ筆者の属性推定を行っている。前述した通り、本研究で定義した経験情報には「誰が」という動作主に関する情報を含めない。

我々は、これまでも経験情報を構造化/知識化するための技術検討を行ってきた[11][12][13]。しかし、これらの手法において抽出していた行動ルールは「状況」と「行動」間のみであり、さらには空間属性を含まなければならないという制約があった。本稿においては、クエリ非依存で収集したブログ記事、いわばブログ空間全体を対象とした大規模解析を行うことで、空間属性を含むという制約なく、あらゆる人間の行動に関するルール発見を行う。また、感情情報抽出を行い、新たなタイプのルールとして、「状況」、「行動」と「主観」との間のルールも抽出する。さらに、相関ルール抽出においては、従来までに用いていた支持度と確信度に加え、リフトや χ^2 値等の様々な指標を用いて、先行研究よりも「興味深い」ルール発見を試みる。

4. 経験マイニング手法の概要

本研究の目的は、CGMテキストに記述された経験情報、いわば“個人の知”そのものの流通を高めることである。この目的に対して、我々は以下の3つのステップが重要だと考えている。

1: 構造化 人間の経験を表現する最小要素 $E = \{ \text{時間, 空間, 動作, 対象, 評価, 感情} \}$ を抽出し、非構造化データであるブログデータを「経験」という観点で構造化する。構造化データを関係データベースに格納すれば、SQLのようなデータベースに対する問い合わせ言語を用いて細かい属性を指定したレコード（経験情報）の検索を行うことが可能である。さらには、SQLが提供する集約演算とソート機能とをうまく組み合わせることで、データを多角的に分析することができる。

2: 知識化 大量の経験情報集合から、人間の行動に関する知識を発見する。具体的には、人間の行動と状況、主観との間の関係をルール形式で抽出する。これまで、データマイニングの分野では、ルールの興味深さを評価する様々な指標が提案され

ている．行動に関する知識と，ルールの興味深さを測る指標との関係性を明らかにし，様々な観点から知識発見を試みる．

3: 情報提示 1と2の処理によって得られた情報を，ユーザのプロファイル情報をもとに効果的に提示する．

本稿においては，この中でも特に1と2のステップについて述べる．5章では1の構造化プロセス，6章では2の知識化プロセスの詳細について述べる．

5. テキストからの経験情報抽出

本章では，評価を除く，以下に示す5要素を抽出し，経験を構造化する手法について述べる．

$E = \{ \text{時間, 空間, 動作, 対象, 感情} \}$

さらに，それぞれの経験情報が，動作主にとって，成功だったのか，それとも失敗だったのかという観点で，成功/失敗要素を付与する．本研究においては「成功/失敗は主に動作主の感情に因る」と仮定する．次節以降で，その抽出手法を述べる．

5.1 時間/空間要素の抽出

主要なブログホスティングサービスにおいては，ブログ記事のRSS配信を行っており，メタデータのひとつとして，ブログ記事が投稿された日時を取得することができる．本手法では，このメタデータから時間要素を抽出する．空間要素は，既存の固有表現抽出技術を用いて抽出を行う．固有表現抽出技術は，入力として与えられた文書中から人名，地名や組織名といった固有表現を抽出する技術である．本手法では，固有表現抽出技術を用いて地名，組織名と判定された語を空間要素として抽出する．

5.2 行動(動作とその対象)要素の抽出

一般に，文はどのような動詞で終わるかで(1)する文(2)なる文(3)である文の3タイプに分類できる．行動文は，このうちの「する文」に該当する表現であり，それを構成するのが「動作」とその「対象」である．この抽出においては，筆者ら[11][12]の先行研究における抽出技術を用いた．この手法は，自然言語処理と辞書照合がベースであり，動作と対象をペアで抽出する．その処理の概要を以下に示す．

- (1) 日本語語彙大系を利用した動作動詞の取得
- (2) 移動を示す動詞の削除
- (3) 深層格における対象格の抽出

1と2の処理によって，動作を抽出し，3の処理によって，抽出した動作の対象を抽出する．なお，2の処理は「行く」「来る」や「訪れる」のような移動を示す動詞を削除する処理である．「行く」は「初詣に行く」のように行動として用いられることもあるが「京都に行く」のように，移動，つまり空間要素を示すのに用いられることもある．今回は，抽出精度を重視し，移動文は削除するという方針をとった．

5.3 感情要素の抽出

経験を構成する主観は，評価と感情とに細分化できる．「評価」は「良い」や「悪い」のような，ある対象に対する主観的な価値付けである．一方で「感情」は「嬉しい」や「悲しい」のような，人間の主体自身に関する心的状態を示している．

本手法では，後者の感情を示す表現を手手で収集し，感情語

表1 登録した感情語の一例

カテゴリ	感情語
喜び (30)	嬉しい, 笑う, 爆笑, 満足, 感動, 満喫
驚き (10)	衝撃, 驚く, 混乱, 動揺, びっくり
困惑 (12)	困る, 悩む, 苦悩, 苦渋, 落胆, 凹む
怒り (6)	怒る, 憤り, 苛立つ, 腹が立つ, 非難
悲しみ (11)	悲しい, 涙, 嘆く, 悲痛, 号泣, 切ない
疲労 (19)	疲れる, 疲労, ぐったり, がっかり
不安 (14)	心配, 気がかり, おびえる, 怖い, 恐い
不満 (19)	不満, 不平, 後悔, 悔しい, つまらない

辞書を構築した．また，すべての感情語を，福原ら[7]，熊本ら[8]，Plutchik[9]らの研究に基づき，喜び，驚き，困惑，怒り，悲しみ，疲労，不安，不満という8カテゴリに分類した．表1に，収集した感情語の一例を示す．括弧内の数値は，それぞれのカテゴリに含まれる感情語数であり，その総数は121である．

5.4 成功/失敗要素の付与

テキストから抽出した経験情報に対して，動作主にとって成功だったのか，それとも失敗だったのかという観点で，成功/失敗要素を付与する．成功が失敗かの判断は，動作主の主観(「評価」と動作主が抱いた「感情」)に因ると考えられるが，特に「感情」に因るところが大きい．以下に示す経験情報の抽出例について考える．

(例) 昨日，清水寺に紅葉を見に行きました．確かに紅葉はきれいだったのですが，観光客で大混雑…後悔しました．

{ 時間, 空間, 動作, 対象, 評価, 感情 } = { 昨日, 清水寺, 見る, 紅葉, きれい, 後悔 }

この例において，紅葉(対象)に対する評価はpositive(きれい)だが，この行動を行った結果，動作主が抱いた感情はnegative(後悔)である．つまり，この動作主は「清水寺に紅葉を見に行く」という行動選択を後悔しており，失敗だったと考えている．評価が対象に対する価値付けであるのに対して，人間の感情は，ある行動に対する価値付けであるといえる．この観点で付与された成功/失敗要素は，他人がその行動を行うべきか否かの行動判断に役立てることができる．

本手法においては，ある行動を起こした結果，動作主が肯定的な感情を持った場合は成功であり，否定的な感情を持った場合は，失敗とみなし，得られた経験情報の感情要素から成功/失敗を導き出す．具体的には，感情カテゴリが「喜び」の感情語に対しては成功，「困惑」，「怒り」，「悲しみ」，「疲労」，「不安」，「不満」に対しては失敗，また「驚き」に関しては，肯定，否定の判断が難しいため，今回は肯定/否定以外とした．例えば，「びっくり」のような感情語は，良い意味と悪い意味の両方で用いられることがあるからである．

5.5 抽出データの関係データベースへの格納

得られた経験情報を関係データベースに格納する．格納する形式を以下に示す．

$R(DID, \text{時間, 空間, 行動(動作, 対象), 感情, 結果})$

DIDは，ブログ記事のIDであり，結果属性には成功/失敗判

定によって導出した値(成功/失敗/それ以外)を格納する。ただし、今回の手法においては、同一記事から空間属性、行動属性、感情属性の要素がそれぞれ複数得られる場合がある。関係データベースは、単純な二次元の表であるので、1つのブログ記事から得られた情報を1レコードとして表現することができない。属性間の要素の正しい組み合わせは、今後取り組む課題とし、今回のアプローチではそれぞれの属性間のすべての組み合わせを生成し、関係データベースに格納する。

5.6 データベースの問い合わせ

前節で説明した手法により、関係データベースへの問い合わせ言語であるSQLを用いて、データを分析することが可能となる。SQLは、細かい属性を指定したレコードの検索に留まらず、部分集合を定義することによる集約演算、複数の集合間の集合演算等を行う機能を備えている。集約演算をうまく利用すれば、ソート機能と組み合わせ、データを多角的に分析することができる。例えば、2007年4月に多くの人々が桜を見た場所を求めるためには、SQLで次のように記述する。

```
SELECT 場所,COUNT(*) AS C FROM TABLE
WHERE 動作='見る' AND 対象='桜' AND (時間 BETWEEN '2007/04/01' AND '2007/04/31')
```

```
GROUP BY 場所 ORDER BY C DESC
```

また、集合を扱うことに長けているのもSQLの特徴である。集合演算は、2つのクエリの検索結果を、和、差、共通部分の3種類の方法で結合する。集合演算を用いることで、複数の検索結果を比較することができる。例えば、京都で必ず失敗する行動(動作, 対象)を求めるためにはSQLで次のように記述する。

```
SELECT 動作,対象 FROM TABLE
WHERE 空間='京都' AND 結果='失敗'
EXCEPT
SELECT 動作,対象 FROM TABLE
WHERE 空間='京都' AND 結果='成功'
```

6. 経験情報集合からの知識発見

前章までの手法で、経験情報に対する構造的なアクセスが可能となったが、本章では、さらに、得られた大量の経験情報集合から、相関ルール抽出技術を用いて、状況、行動、主観との間に存在する関係性をルール形式で抽出する。相関ルールは、 $A \rightarrow B$ という形式で表現され、これは「Aが起こったという前提のもとで、Bも同時に起こる」ことを示す。また、Aをルールの条件部、Bを結論部と呼ぶ。具体的には、以下に示す4タイプのルールを抽出する。

Type 1: 状況と行動 : [空間, 時間] \rightarrow [動作, 対象]

(例1) 北海道, 5月 \rightarrow [時間] 見る, 桜 [動作, 対象]

Type 2: 状況と主観 : [空間, 時間] \rightarrow [感情]

(例2) ディズニーランド [空間] \rightarrow 楽しい [感情]

Type 3: 行動と主観 : [動作, 対象] \rightarrow [感情]

(例3) 引く, おみくじ [動作, 対象] \rightarrow がっかり [感情]

Type 4: 状況と行動と主観 :

[時間, 空間, 動作, 対象] \rightarrow [感情]

[時間, 空間] \rightarrow [動作, 対象, 感情]

[動作, 対象] \rightarrow [時間, 空間, 感情]

(例4) 貰う, 義理チョコ [動作, 対象] \rightarrow 困る [感情]

なお、上記の4タイプのルールにおいて、結論部と条件部が入れ替わったルールも抽出対象とする。従来までの著者らの取り組みにおいては、主観を含まないタイプ1のルールのみを抽出していたが、本稿では新たに主観を含む残り3タイプのルールを抽出する。知識発見分野においては、相関ルールの興味深さを測る様々な指標が提案されている。それぞれの指標は、データの異なる側面を評価するため、得られる結果も異なり、多様な観点からのデータ分析が可能である。次節では、データベースからの知識発見分野におけるルールの興味深さの指標について説明する。

6.1 ルールの興味深さの指標

ルールの興味深さを測る指標は、客観的指標 (*objective measure*) と主観的指標 (*subjective measure*) とに大別できる。客観的指標が、データのみ依存する指標である一方、主観的指標は、データ自身に加え、ユーザの知識や背景をも考慮する。客観的指標によるルールの評価において、最も重視されているのは、ルールの一般性 (*generality*) と、信頼性 (*reliability*) という側面である。一般性とは、データの特徴をどの程度反映しているかという観点でルールを評価するものであり、支持度 (*support*) や被覆度 (*coverage*) などがこれに該当する。全レコード数(ブログ記事数)を N 、集合 A を含むレコード数(ブログ記事数)を $n(A)$ 、集合 A と集合 B をともに含むレコード数(ブログ記事数)を $n(AB)$ としたとき、ルール $A \rightarrow B$ の支持度は $P(AB) = \frac{n(AB)}{N}$ で、被覆度は $P(A) = \frac{n(A)}{N}$ で表される。信頼性の評価には、確信度 (*confidence*) やリフト (*lift*) などの指標が用いられる。確信度は、集合 A が与えられたときの条件付き確率 $P(B|A) = \frac{P(AB)}{P(A)}$ で表され、条件付き確率が大きいほど信頼性の高いルールとする。しかし、確信度には欠点がある。例えば、ルール $A \rightarrow B$ について確信度が75%だった場合を考える。これは、非常に高い数値であるが、そもそも集合 B の出現確率 $P(B)$ が80%だったら、むしろ、結論部の予測に、条件 A がマイナスに働いていることになる。他のルールとの相対性に基づいてルールを評価する指標が、リフトである。リフトは、 $\frac{P(B|A)}{P(B)}$ で定義され、直感的には、条件 A を加えることで B が何倍起こりやすくなるかを示しているといえる。ここに示した尺度以外にも、統計に基づく指標である χ^2 値 [3] や、情報量に基づく指標である J-measure [4] など、統計学、情報理論、情報検索などの分野に起因し、様々な客観的な指標が提案されている。それぞれの指標の特性を理解し、効果的に利用することで、経験情報集合の中から様々な知識が発見できるといえる。

8章の評価実験においては、データのみ依存する客観的指標に属するこれらの指標と、得られる行動ルールの関係について、実データを用いて分析する。詳細は後述するが、確信度と支持度がある程度高いルールを取得し、そのうちリフトが高いものから吟味していくというアプローチが有効であることがわかった。次章では、そのアプローチを実装したプロトタイプシステムについて述べる。

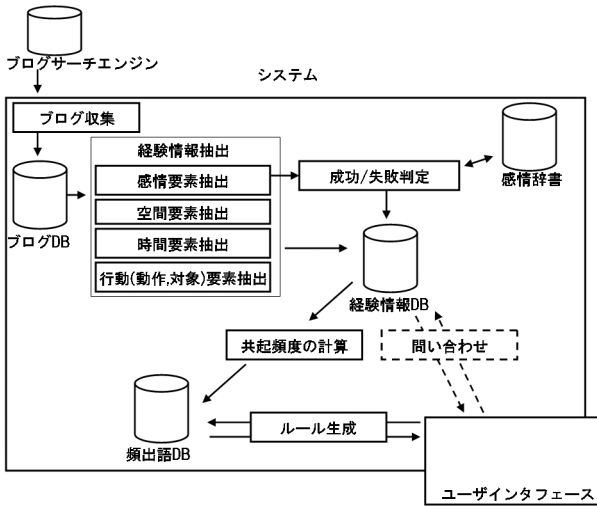


図 2 システムの構成

7. プロトタイプシステム

前章までに説明した手法に基づき、システムを実装した。ユーザは、経験情報データベースから得られたルールを検索することができる。「時間」「空間」「動作」「対象」「感情」「結果」属性の要素を検索条件に指定すると、システムは、それらを条件部に持つルールを生成し、その結論部の要素（指定していない属性の要素）を確信度の降順で提示する。また、ユーザは、ルールの価値を多面的に評価することもできる。

7.1 システムの構成

プロトタイプシステムの構成を図 2 に示す。ブログの収集には、BLOGRANGER 2.0 API [14] を利用した。ブログ文書の形態素解析器には日本語形態素解析ソフト JTAG [15]、空間要素の抽出には、多言語固有表現抽出器 Namelister [16] を使用した。本システムにおいては、ブログの収集、経験情報の抽出、経験情報の成功/失敗分類と経験情報データベースへの格納をバッチ処理にて行う。使用したデータベースは MySQL [17] である。また、相関ルールの抽出には、Apriori アルゴリズム [18] を用いた。処理を構成する (1) 共起頻度の計算 (2) ルールの生成のうち、計算コストが高い (1) の処理のみをバッチ処理にて行う。(2) の処理は、ユーザとのインタラクションに応じて、リアルタイムで行う。次節で、システムのユーザ操作について述べる。

7.2 ユーザ操作

システムのユーザインタフェースを図 3 に示す。システムの基本的な操作の流れは以下の通りである。

(1) ユーザ: 検索条件と出力条件を指定する。検索条件に入力できるのは、時間、空間、動作、対象、感情、結果属性の要素、もしくはその組み合わせであり、出力条件には、検索条件と重複しない属性を指定する。また、出力件数 N を指定する。

(2) システム: 検索条件に指定した要素の集合を条件部に、結論部に出力条件の属性要素を持つルールを生成し、確信度の降順に上位 N 件をユーザに提示する。それと同時に、上位 N 件のルールについてリフトを計算する。計算したリフトは、インタフェース上に提示されたルールのフォントに反映される。

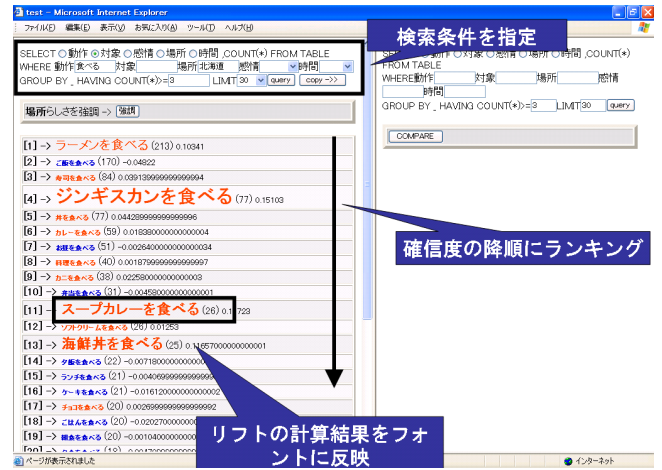


図 3 システムのユーザインタフェース

リフトが 1 以上のルールのフォントの色は赤、1 未満のものは青であり、フォントのサイズはリフトの絶対値を表現する。

(3) ユーザ: 得られたルールを再評価する属性（時間、空間、行動（動作と対象）、感情、結果）を指定する。

(4) システム: ユーザが指定した属性に基づき、上位 N 件のルールを再評価する。この結果も 2 と同様、フォントで表現する。

例えば、検索条件の動作属性の要素に「買う」、結果属性の要素に「失敗」を入力し、出力属性に対象を指定すれば、「買って失敗したもの」を検索することができる。次節で、ルールを再評価する操作 3,4 の詳細を述べる。

7.3 ルールの多面的な評価

ユーザが条件部に複数の属性要素を指定した場合、条件部のどの要素が、結論部の予測に寄与しているかを分析することができる。インタフェース上には、「時間」、「空間」、「行動（動作、対象）」、「感情」、「結果」と書かれた 5 種類のボタンが配置されており、ボタンを押すことで、その属性の影響を確認することができる。例えば、以下のルールが検索結果として得られた場合を考える。

(a) 京都駅, 4月 → バス, 乗る (確信度:70%)

(a) のルールは、「京都駅に 4 月に行ったという経験のうちの 70% は、バスに乗っている」ことを意味する。この値は確信度としては高い値であり、これのみで有益な情報と捉えることもできるが、それと類似した以下のルールと比較をすればさらに情報を引き出すことができる。

(b) 4月 → バス, 乗る (確信度:80%)

(c) 京都駅 → バス, 乗る (確信度:60%)

(a) の確信度 70% は、(b) の確信度 80% と比べれば低い数値であり、これはつまり「4 月においては、京都駅でバスに乗る人が他の場所に比べて少ない」ことを意味する。また、(c) と比べることで、「京都駅においては、4 月にバスに乗る人が他の時期に比べて多い」ことがわかる。前者は、(a) のルールを空間という側面で評価し、後者は、時間という側面で評価しているといえる。 $A \rightarrow B$ の条件部 A の部分集合 $C \subset A$ が集合 B の予測にどの程度寄与しているかを調べる場合、システムは、集合 A からその部分集合 C を除いた集合 A' に関するルール

表 2 解析したブログ記事と抽出データに関する情報

記事数	48,112,100
収集期間	2007/1/1 ~ 5/31
1以上の要素が抽出できた記事数	29,778,231
1記事当たりの平均抽出要素数	4.811
1記事当たりの平均抽出要素数(行動)	3.777
1記事当たりの平均抽出要素数(空間)	0.845
1記事当たりの平均抽出要素数(感情)	0.259
空間属性のユニーク要素数	34,932
行動属性のユニーク要素数	664,914
感情属性のユニーク要素数	121

$A' \rightarrow B$ を取得し, $\frac{P(B|A)}{P(B|A')}$ を評価値として得る. この考え方は, 他のルールとの相対性に基づいてルールを評価するという点で, リフトの拡張と捉えることができる.

8. 評価実験

本章では, データマイニング分野におけるルールの客観的指標と, それによって得られるルールの特性との関係性を議論する.

8.1 関連ルールの前処理

関連ルールを導出する前処理として, ブログ記事から実際に経験情報を抽出し, その結果集合から要素間の共起数を算出した. なお, 今回の実験では, 計算量を減らすために動作と対象を行動という1つの属性として扱い, 時間, 空間, 行動, 感情という4属性間の共起頻度のみを抽出する. 時間属性の要素は日付情報は連続値であるため, ブログを5日間づつ60区間に分け, それぞれの区間で, 空間, 行動, 感情という3属性間の共起頻度を算出した. 全区間トータルでの正確な共起頻度を算出することは膨大な計算量が必要となるため, 各区間単独で, 共起頻度が3以上の組み合わせを取得し, それらを集計することで計算量を削減した. 解析を行ったブログ記事と抽出した経験情報の詳細を表2に示す. クローラーで収集した5ヶ月間分のブログデータすべてから経験情報を抽出した. 解析したブログ総数は約4800万件である. なお, 1記事当たりの平均抽出要素数は, 時間属性を除くものであり, これを含めると, 平均で5.811ということになる.

8.2 導出される行動ルールと客観的指標との関係性の分析

最初に, 支持度, 確信度, リフトという3つの指標で得られるルールについて考察する. 人間の「主観」を含むタイプ2からタイプ4のルールは, 定量的な評価が難しいため, タイプ1に該当する以下の形式のルールを用いて, それぞれの指標で得られる結果を比較する. なお, ルール抽出に設けた制約条件も同時に示す.

(1) 空間 → 動作, 対象

- 最小支持度: 1.00E-07
- 最小確信度: 1.00E-04
- 空間属性: お台場, ディズニーランド, 横浜, 沖縄, 京都, 大阪, 北海道, 名古屋 (8)
- 動作属性: 食べる, 見る, 買う (3)

- ルール総数: 1659

(2) 時間 → 動作, 対象

- 最小支持度: 1.00E-05
- 最小確信度: 1.00E-04
- 時間属性: 2007年1月, 2月, 3月, 4月, 5月 (5)
- 動作属性: 食べる, 見る, 買う (3)
- ルール総数: 1330

次に得られたルール集合を支持度, 確信度, リフトの降順にソートする. 上記の(1)において, 支持度, 確信度, リフトでソートした場合の上位の結果を表3に示す.

支持度はルールの条件部と結論部の単純な共起頻度に基づく評価尺度であるため, 「京都 → 桜を見る」や「大阪 → ご飯を食べる」のような, 一般的な組み合わせを評価する傾向にある. つまり, そもそも条件部と結論部の語の出現頻度が高い組み合わせを評価する. 確信度はルールとしての信頼性を評価する尺度であり, 条件部を満たした場合の結論部の生起確率を示す. つまり, 確信度による評価値の高いルール「ディズニーランド → パレードを見る」は, ディズニーランドに行く人の多くはパレードを見ていることを意味する. しかし, 上位の結果からわかるように, 確信度に基づいて得られるルールは, 多くの人々にとって既知の事実であり, 興味深さに欠ける. その原因は「ご飯を食べる」のような, 結論部(行動語)の出現確率がそもそも高いルールを評価していることにある. 前述した通り, 条件部を加えたときの結論部の出現確率が, 全体集合におけるそもそもの出現確率と比較して, どの程度上昇しているかを評価する指標がリフトである. 実際, リフトによる評価では「お台場 → ご飯を食べる」や「横浜 → ご飯を食べる」の評価値は低くなっている. リフトによるソートの上位の結果を見ると, 結論部(行動語)に「スプリングロールを食べる」や「シンデレラ城を見る」等の出現確率がそれほど高くない行動に関するルールを評価していることがわかる. つまり, リフトを計算することにより, 一般的な行動に関するルールを除去し, ある空間(時間)に特有な行動を示すルールのみを抽出することができる.

次に, ある空間(時間)に特有な行動を得るための指標としてのリフトの有用性を定量的に検証する. 比較対象とする指標は, 支持度, 確信度に加え, χ^2 値 [3], J-measure [4] の4つである. J-measure は, 情報量に基づく指標であり, $P(AB) * \log\left(\frac{P(B|A)}{P(B)}\right) + P(A \neg B) * \log\left(\frac{P(\neg B|A)}{P(\neg B)}\right)$ で計算する. この式から, J-measure は, 支持度 $P(AB)$, 確信度 $P(B|A)$, リフト $\frac{P(B|A)}{P(B)}$ を統合した指標であることがわかる. (1)においては「時間に特有な対象」, (2)においては「空間に特有な対象」がルールの結論部として得られた場合に正解とする. 正解セットは, 人手で作成した. 作成に関わった人数は2人であり, 今回は, 両者の意見が一致したもののみを正解としている. (1)の正解数は129(ルール総数は1659)であり, (2)の正解数は55(ルール総数は1330)である. それぞれの指標でルールをソートし, (1)においては上位129件の, (2)においては上位55件の適合率を評価した. 適合率は以下に示す指標である.

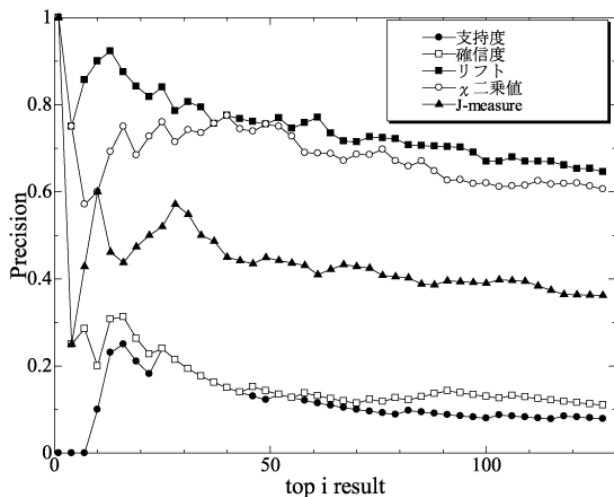


図 4 形式 (1) におけるルールの評価指標と適合率との関係

$$\text{適合率} = \frac{\text{適合ルール数}}{\text{総ルール数}}$$

表 4 にその結果を示す。また、図 4 に、上位 i 件までの適合率の変化を示す。実験の結果、リフトがその他の指標よりも高精度に空間（時間）に特有な行動を抽出していることがわかる。特に、支持度、及び確信度との比較においては、大きな精度改善が見られる。また、 χ^2 値によるルールの評価は、リフトに次ぐ精度が出た。 χ^2 値は、条件部と結論部の方向性を考慮せず、その組み合わせの依存性を測る尺度である。「状況 → 行動」ルールを評価する本実験においては、ルールの recall も重要な指標である傾向であったため、同程度の精度が出たと考えることができる。recall とは、 $P(A|B) = \frac{P(AB)}{P(B)}$ で計算でき、条件部と結論部を入れ替えたルール、つまり「ディズニーランド → スプリングロール」に対する「スプリングロール → ディズニーランド」の確信度に等しくなる。J-measure は、3 番目に高い精度であった。J-measure は、支持度と確信度とリフトを統合した指標であるが、支持度と確信度による精度が低かったため、これらの影響が精度低下につながったと考えられる。

今回の実験では、支持度と確信度に加え、リフトを計算することで行動に関する興味深いルールが得られることがわかった。次節では、同様のアプローチを用いて、タイプ 2 からタイプ 4 の、人間の「主観」に関するルールを抽出した結果について述べる。

8.3 主観に関して得られたルールの考察

主観に関して得られたルールの一例を表 5、表 6 に示す。表 5 は、[空間]→[感情] という形式のルール、表 6 はその条件部に時間属性が加わった [時間, 空間]→[感情] という形式のルールであり、いずれも空間属性の要素が「ディズニーランド」である。表 5 を見る限り、一般的に、ディズニーランドは多くの人にとって「喜び」をもたらす場所であることがわかる。一方、表 6 を見ると、5 月、3 月に訪れた場合には、「怒り」や「不満」が多く失敗の傾向が強いことがわかる。こういった失敗の背景には、春休み（3 月）やゴールデンウィーク（5 月）による園内の混雑が原因であると考えられる。また、この情報は、一般的な知識に対する例外と捉えることができ、常識

表 3 支持度、確信度、リフトを用いて形式 (1) のルールをソートした場合の上位 10 件の結果

支持度		支持度	確信度	リフト
1	京都 → 桜を見る	2.06E-05	9.79E-03	10.26
2	大阪 → ご飯を食べる	2.04E-05	7.71E-03	2.06
3	京都 → ご飯を食べる	1.48E-05	7.02E-03	1.88
4	沖縄 → 料理を食べる	1.05E-05	8.19E-03	14.30
5	横浜 → ご飯を食べる	9.97E-06	9.98E-03	2.67
6	沖縄 → そばを食べる	9.60E-06	7.49E-03	39.64
7	大阪 → 顔を見る	8.66E-06	3.27E-03	1.61
8	大阪 → 姿を見る	7.99E-06	3.02E-03	1.36
9	京都 → お土産を買う	7.99E-06	3.80E-03	10.30
10	ディズニーランド → パレードを見る	7.99E-06	1.23E-02	262.59

確信度		支持度	確信度	リフト
1	ディズニーランド → パレードを見る	7.99E-06	1.23E-02	262.59
2	お台場 → ご飯を食べる	5.27E-06	1.08E-02	2.87
3	横浜 → ご飯を食べる	9.97E-06	9.98E-03	2.67
4	京都 → 桜を見る	2.06E-05	9.79E-03	10.26
5	ディズニーランド → ショーを見る	5.57E-06	8.55E-03	36.21
6	ディズニーランド → ご飯を食べる	5.37E-06	8.24E-03	2.20
7	沖縄 → 料理を食べる	1.05E-05	8.19E-03	14.30
8	大阪 → ご飯を食べる	2.04E-05	7.71E-03	2.06
9	沖縄 → そばを食べる	9.60E-06	7.49E-03	39.64
10	名古屋 → ご飯を食べる	7.05E-06	7.25E-03	1.94

リフト		支持度	確信度	リフト
1	ディズニーランド → スプリングロールを食べる	1.01E-07	1.55E-04	1533.62
2	ディズニーランド → シンデレラ城を見る	2.01E-07	3.09E-04	1533.62
3	ディズニーランド → エレクトリカルパレードを見る	1.01E-07	1.55E-04	1150.21
4	お台場 → ベール展を見る	1.34E-07	2.74E-04	815.84
5	お台場 → コルベール展を見る	1.34E-07	2.74E-04	815.84
6	ディズニーランド → ステイッチのパレードを見る	3.36E-07	5.15E-04	766.81
7	お台場 → インポートカーショーを見る	1.01E-07	2.05E-04	764.85
8	お台場 → グレゴリーコルベール展を見る	1.01E-07	2.05E-04	679.87
9	お台場 → snowを見る	2.35E-07	4.79E-04	648.97
10	お台場 → カーショーを見る	1.01E-07	2.05E-04	556.26

表 4 ルールの評価尺度と適合率との関係

	支持度	確信度	リフト	χ^2 値	J-measure
(1)	0.085	0.109	0.643	0.596	0.357
(2)	0.018	0.018	0.727	0.618	0.491

を覆すルールとして価値があるといえる。表7に、行動と主観との関係を表現するタイプ3について得られた結果の一例も示す。「Wiiを買うのは喜び」や「マンションを買うのは不安」といった人間の購買行動と主観との関係を表している。一方、タイプ4のルールで、空間、行動、感情の3属性をとともに含むものに関しては、一部の有名な場所に関するルールを除き、多くのルールを発見することができなかった。これは、3属性(空間、行動、感情)要素のすべてを1記事中に含むものがそもそも少ないためである(2007年1月1日~1月5日までの期間で、3記事以上に出現した3属性間の組み合わせは1362種類であり、2属性間の組み合わせの259259種類と比べて少なかった)。我々は、抽出期間を広げ、解析データの量を増やすことでこの問題は解決可能だと考えている。

表5 ルール形式:[空間]→[感情] (空間="ディズニーランド")

ルール	リフト	確信度
ディズニーランド → 喜び	1.166	0.583
ディズニーランド → 疲労	0.916	0.117
ディズニーランド → 不満	0.906	0.113
ディズニーランド → 不安	0.866	0.068

表6 ルール形式:[時間][空間]→[感情] (空間="ディズニーランド", 時間="1月","2月","3月","4月","5月")

結果	ルール	リフト	確信度
失敗	5月, ディズニーランド → 怒り	1.146	0.017
	3月, ディズニーランド → 悲しみ	1.191	0.051
	1月, ディズニーランド → 不満	1.112	0.126
成功	2月, ディズニーランド → 喜び	1.023	0.600
	4月, ディズニーランド → 喜び	1.013	0.591

表7 ルール形式:[時間][動作][感情]→[対象] (時間="1月", 動作="買う", 感情="喜び","疲労","不安")

ルール	リフト	確信度
1月, 喜び → おみやげを買う	1.559	4.126E-05
1月, 喜び → 指輪を買う	1.400	4.413E-05
1月, 喜び → クリスマスプレゼントを買う	1.238	4.700E-05
1月, 喜び → 券を買う	1.185	6.709E-05
1月, 喜び → Wii を買う	1.134	5.274E-05
1月, 疲労 → 風邪薬を買う	6.375	5.624E-05
1月, 疲労 → 枕を買う	3.749	4.614E-05
1月, 疲労 → 加湿器を買う	3.671	6.777E-05
1月, 疲労 → 器を買う	2.530	8.652E-05
1月, 疲労 → 飴を買う	2.412	5.624E-05
1月, 不安 → お酒を買う	1.701	4.442E-05
1月, 不安 → マンションを買う	1.623	3.507E-05
1月, 不安 → おみやげを買う	1.590	4.208E-05
1月, 不安 → 靴下を買う	1.568	3.741E-05
1月, 不安 → コンタクトを買う	1.390	3.507E-05

9. まとめと今後の課題

本稿においては、ブログに記述された人間の経験を構造化する試みとして、経験を構成する5要素(時間、空間、動作、対象、感情)をブログから抽出し、さらに「成功/失敗は主に動作主の感情に因る」として、それぞれの経験情報に成功/失敗要素を付与する手法を示した。また、経験を知識化する試みとして、約4800万件のブログ記事から抽出した大量の経験情報集合から、相関ルール抽出技術を用いて、状況と行動と主観との関係をルール形式で抽出した。ルールの評価指標としては、特にリフトが、興味深いルールの抽出に効果的に働くことを示した。今後は、本手法によって構造化/知識化した人間の経験を、ユーザに提示する手法について検討を行う予定である。

文 献

- [1] 奥村 学, 南野 朋之, 藤木稔明, 鈴木泰裕, "Automatic Collection and Monitoring of Japanese Weblogs", FIT2004 7K-6, 2004.
- [2] kizasi.jp <http://kizasi.jp/>
- [3] 福田剛志, 森下真一, "相関ルールの可視化について", 電子情報通信学会技術研究報告, 95-81, pp.41-48.
- [4] P.Smyth and M.Goodman, "Rule Induction using Information Theory", Knowledge Discovery in Databases, pp. 159-176, AAAI/MIT Press, 1991.
- [5] 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", 情報処理学会自然言語処理研究会 (NL-144-11).
- [6] Bing Liu, Mingqing Hu and Junsheng Cheng. "Opinion Observer: Analyzing and Comparing Opinions on the Web" In Proceedings of the 14th international World Wide Web conference (WWW-2005), May 10-14, 2005.
- [7] 福原知宏, 中川裕志, 西田豊明, "感情表現と用語のクラスタリングを用いた時系列テキスト集合からの話題検出", 人工知能学会全国大会, 2E1-02 2006.
- [8] 熊本忠彦, 田中克己, "Web ニュース記事を対象とする喜怒哀楽抽出システム", 情報処理学会研究報告, NL165-3, pp.2003.
- [9] Plutchik, R. The Multifactor-Analytic Theory of Emotion, The Journal of Psychology, Vol. 50, pp. 153-171 1960.
- [10] 安田宜仁, 平尾努, 鈴木潤, 磯崎秀樹, "ブログ作者の居住域の推定", 自然言語処理学会第12回年次大会 (NLP2006), 2006.
- [11] Takeshi Kurashima, Taro Tezuka, and Katsumi Tanaka, "Blog Map of Experiences: Extracting and Geographically Mapping Visitor Experiences from City Blogs", Web Information Systems (WISE2005), pp. 496-503, November 2005.
- [12] 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), pp. 213-222, Krakow, September 2006.
- [13] Taro Tezuka, Takeshi Kurashima, and Katsumi Tanaka, "Toward Tighter Integration of Web Search with a Geographic Information System", Proceedings of the Fifteenth World Wide Web Conference (WWW2006), pp. 141-144, 2006.
- [14] BLOGRANGER 2.0 API, <http://ranger.labs.goo.ne.jp/TG/webapi.php>
- [15] T. Fuchi, and S. Takagi, "Japanese Morphological Analyzer using Word Co-occurrence-JTAG", COLING-ACL, pp.409-413, 1998.
- [16] 齋藤邦子, 鈴木潤, 今村賢治, "CRFを用いたブログからの固有表現抽出", 言語処理学会第13回年次大会, 2007.
- [17] MySQL, <http://www.mysql.com/>
- [18] R. Agrawal and R. Srikant, Fast algorithms for mining association rules in large databases, Proceedings of the 20th Intl. Conf. on Very Large Data Bases, pp. 487-499, 1994.