2D2-2

Webから見る実世界

Real World Sensing based on Blog Article Analysis

山田 剛一[*1*3] Koichi Yamada　圓戸 辰郎[*2*3] Tatsuro Endo　小林 聡[*1] Satoru Kobayashi　絹川 博之[*1] Hiroshi Kinukawa　田村 陽介[*2*3] Yosuke Tamura

[*1]東京電機大学 未来科学部 School of Science and Technology for Future Life, Tokyo Denki University

[*2]株式会社フィックスターズ Fixstars Corporation

[*3]JST-CREST

On the web, there is so much information of the real world. Everyday and every time, bloggers from all over the world are writing many things that they are watching, hearing, smelling, etc. We collect information of the real world on the blog articles and are going to develop a real world sensing system. In this article, we describe some features of the real world information on the blog articles and propose a method that extracts expressions based on experiences in the real world.

1. はじめに

いま、Web には実世界の情報があふれている。blog や SNS の普及により日記を書く人の数が爆発的に増加し、世界中の多くの人々が、日々、その日見たもの感じたものを言葉として表現し、公開している。

Web から見る実世界と言っても、Web カメラを設置するなど、物理センサを用いるのではない。本研究では、物理センサを用いず、代わりに人間の五感を介して実世界をセンシングする、Web からの実世界センシングを目指している。

Web 上の日記に現れる実世界情報は、人間という知的なセンサが認識・解釈を行った結果であり、Web カメラのような単純な物理デバイスから得られる情報よりも、はるかに抽象度の高い情報となっている。日々更新されるこれらの情報を収集することにより、これまでとは量も質も異なる実世界のセンシングが実現できる。

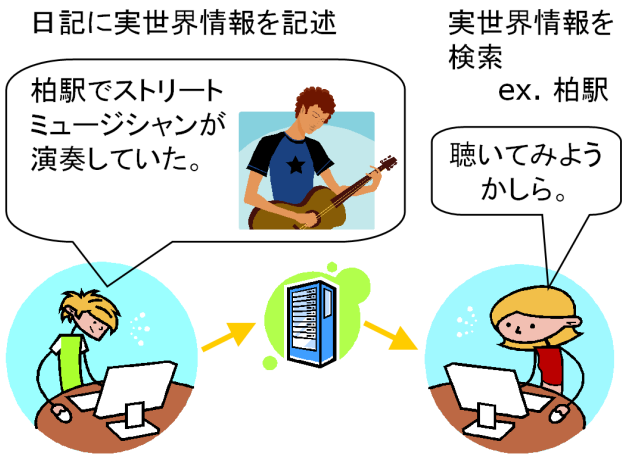

図 1: 日記に現れる実世界情報を検索

連絡先: 山田 剛一.
東京電機大学 未来科学部 情報メディア学科.
東京都千代田区神田錦町 2-2.
ky@acm.org

本論文の前半では、日記の書き手である人間を実世界センサと見立てた場合に、どのような実世界情報を得ることになるのか考察をする。後半では、我々が現在開発している、blog 記事が書き手の実体験に基づく記事であるかを判定するシステムについて述べる。

2. 本研究で扱う実世界情報

実世界情報と言っても、今食べている野菜の残留農薬から、ここ数日の政界再編の動きまで、さまざまである。本研究では、ある「場」において人間が知覚することができる状況を、実世界情報として扱う。

例えば、現在、国立新美術館 (六本木) において「モディリアーニ展」が開催されている[*1]。これは行くべきだろうか? 行くと何が見られるのだろう。混んでいるのだろうか。すでに行った人の感想はどんなものだろう。これらは、すべて blog に書かれている実世界情報である。

単に blog をキーワード検索しただけでは、実際に自ら体験せずに書かれている記事、実際に行っていてもその場に関する情報がない記事、スパムブログなどに阻まれ、目的とする実世界情報に到達しにくい。本研究では、実際に体験した上での記述であるかを言語表現により判断し、その場で人間が知覚した対象についての情報を抽出する。

blog 記事の分析の研究は、評価情報 (評判情報) を扱うものが主流である。抽出項目は評価対象・属性・評価値の 3 項であり、書き手がその評価をどのような場で行ったのか、という情報は扱われない。一方、評価に限らず、経験一般を対象とする研究も行われている [池田 07, 倉島 08, 乾 08]。これらの研究で扱われる表現には実世界情報を表すものも多く含まれるが、これらは経験情報を抽出する一般的な枠組みであり、本研究で扱っているような実世界情報に特化した解析は行わない。本研究では実世界情報に的を絞り、ある特定の場で、書き手が実際に体験して記述した情報を扱う。

3. 実世界センサとしての人間の特性

人間、特に日記の書き手を実世界センサに見立てたとき、センサとしての特性はどのようなものだろうか。外界の対象を

*1　2008 年 3 月 26 日 (水)—6 月 9 日 (月)

認識してから言葉として表現するまでの過程は単純ではなく、通常の物理的なセンサとは特性が大きく異なる。

3.1 人間は対象を適当に扱う

物理的なセンサと比較すると、入り口の段階ですでに大きな差がある。人間は認識すべき対象を、適度な抽象度で認識することができる。それだけではなく、誤認することもあれば、見なかったことにすることもある。

実世界センサとして人間を利用しようという立場からすると、その稼働率の低さは特筆すべきものがある。人間は起きている間、常に何かしら知覚しているが、それが記述の対象となることは稀である。日記に現れてくるものは、その人にとっての非日常的なイベントであることが多く、たとえ blog を常時監視していたとしてもその人の日常をセンシングしていることにはならない。ただ「モディリアーニ展」の例のように、抽出すべき有用な情報は非日常的な情報であることが多いため、うまく情報がフィルタリングされていると捉えることもできる。

3.2 人間は勝手に動く

設置場所が固定のセンサと異なり、人間は主体的に場所を移動する。よって、人間が五感を用いてセンスした値は、日時と場所の情報を含めて認識する必要がある。

日記に現れる実世界情報は、次の 2 つに分類できる。

センサが知覚した実世界 書き手にとっての外界の情報であり、実世界センサとしてのセンシング値。

実世界の中にいるセンサ自身 書き手自身の目から記述される、書き手の行動情報・体験情報など。

本研究では人間をセンサとして見立てているのであるから、そのセンシング値として前者の情報が必要である。一方で、センサの位置やセンシングの状況などは後者の情報、つまり書き手の行動情報によって得られるものであるので、どちらの情報についても分析が必要となる。

3.3 実世界情報と行動・体験・評価情報

すでに若干述べたように、我々が扱う実世界情報と、行動情報・体験情報・評価情報については、いろいろな側面で関連がある。これらについては多くの研究 (例えば [池田 07] [倉島 05] [倉島 08] [乾 08]) が行われているので、これらの情報と実世界情報がどのような関係となっているのかを概観しておく。

行動情報

人間の「行動」は、広く捉えれば体験や評価するという行為までを含むが、それらについては別に述べるのでここでは含めずに考える。それでも、主語が一人称で述語が動詞であれば、なんらかの書き手の行動を表していることが多く、「秋葉原に行く」から「目を凝らす」まで、多種多様である。

実世界情報という観点では、書き手の実世界での有り様という意味においては書き手のすべての行動情報は実世界情報であるが、書き手を人間センサとして見立てたときには、センサの位置やその変化を表す表現が重要である。例えば、書き手の場の移動を示す「行く」「来る」「帰る」などの語がそれにあたる。

体験情報

人間の「体験」にはいくつかタイプがあり、それらは表現にも違いがある。書き手を人間センサとして見立てる立場からすると、人間の五感による知覚を直接表す表現「見る」「聞く」「触れる」「味わう」「嗅ぐ」などは、センシングの行為を直接的に示しており、同時にセンシングの内容も記述されるため重要である。

(1) 武道館でビートルズのライブを見た。

この例の場合、センシング値「ビートルズのライブ」だけでなく、センサの位置「武道館」も示されている。このように場が明示されている場合には、書き手の行動 (場所の移動) を追わなくても場が特定できる。

なお、直接センシングを示す語がなくても、五感による知覚を前提としている表現は多い。

(2) 今日本郷に行ったら、赤門にハトがとまっていた。

この例の場合、「ハトがとまっていた」のを「見た」のであるが、それは明示的に記述されていない。

五感による知覚とは異なる意味として、行動一般による「体験」がある。試行的な行動の場合には「〜てみる」という表現が使われるなど特徴もあるが、人間センサとしての観点からは、行動が「体験」として捉えられるか否か、あるいは試行的であるか否かによってセンサの位置が変化するわけではないので、一般の行動情報と同様に扱ってよいと考えられる。

受動的な体験については、書き手に影響を及ぼした対象を、センシングの対象と考えるかどうかによって状況が異なってくる。

(3) 渋谷を歩いていたら警察官に呼び止められた。

これは、渋谷に人を呼び止める警察官がいた、という実世界情報と捉えることができる。

実世界情報を抽出する観点からは、まず記事の記述が体験に基づくものなのか否かを判別する必要があり、体験情報の記述の有無はその判断の重要な手掛かりとなる。

評価情報

blog などから評価情報 (評判情報) を抽出する研究は、マーケティング等に直結するためさかんに行われている。

評価には、体験に基づいた評価と、そうでない評価がある。また、評価対象が実世界に存在するものと、そうでないものがある。

(4) この旅行のプランはきつすぎる。

(5) ギターの演奏が最高だった。

本研究で扱う実世界情報は与えてくれるのは、後者である。

体験に基づく評価表現があれば、それは実世界に存在する対象を評価している可能性が高い。これは、blog 記事全体が実世界の体験に基づいて書かれているか否かの判断材料となる。

4. Blog からのイベント体験情報抽出システム

Web から実世界を見る、その見方についてはさまざまな形態が考えられるが、ここではその 1 形態として、場を指定してその場の実世界情報を検索する、実世界情報検索システムを考える。

書き手の体験に基づく実世界情報が現れるのは、これまで見てきたような日記タイプのコンテンツである。本研究では日記タイプの blog を対象とし、書き手の実体験に基づいた blog 記事だけを抽出する、blog 検索エンジンを検討している。特

に我々は、場を指定する検索質問としてイベント名 (展覧会やお祭り等の名称) を取り上げ、Blog からのイベント体験情報抽出システムを開発している [小林 07, 小林 08]。

これは blog 検索エンジンのラッパーとして動作し、検索要求であるイベント名を含む blog 記事の中から、書き手の実体験に基づいて記述されている記事だけを抽出し、検索結果として返すものである。このシステムにより、イベントに行くと何が見られるのか、といった実世界情報が含まれている記事のみを得ることができる。

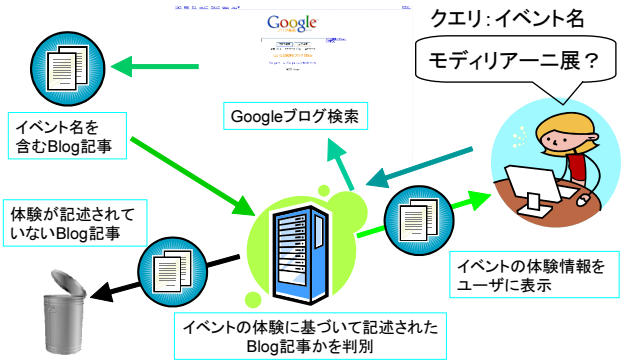

図 2: Blog からのイベント体験情報抽出システム

本システムの処理の流れを以下に示す。

1. イベント名を検索質問として blog 記事を得る
2. 記事タイトル・本文に対して形態素解析・係り受け解析を行う
3. 記事がイベントの実体験に基づいて書かれたものであるかを判定する
4. イベントの実体験に基づいて書かれた記事だけをユーザに提示する

なお、形態素解析には茶筌 [茶筌]、係り受け解析には [工藤 02, 南瓜] を使用している。

4.1 実体験に基づいて記述された記事かの判定

以下のいずれかの条件を満たしている blog 記事は、実体験に基づいて記述された記事であると判定する。

1. 「～に行った」「～を見た」等の「その場での体験を示す表現」が記述されていて、その語に検索質問のイベント名が係っている
2. 「実際にイベントを体験しなければ記述できない表現」が記述されている

4.1.1 「その場での体験を示す表現」の有無による判定

Blog 記事が実際にイベントを体験した上で記述されているのかを判断するために、そのイベントが行われた場所に書き手が存在したことを示す表現に着目する。そのような表現である条件は以下の通りである。

- 「その場での体験を示す表現」が存在する
- 検索質問のイベント名が、その語に係っている

実際には「～に行った」「～を見た」等の、イベント名を伴い、そのイベントが行われた場所へ書き手が現れたことを示す語を辞書として持っている。また、そのような語が記述されていた場合、その語に係っている語をイベント名の候補として抽出し、検索質問のイベント名との照合を行う。

このとき、記事中のイベント名候補と検索質問のイベント名は、完全に一致するとは限らない。表記の揺れや、長いイベント名の省略、通称表現の利用などがその原因である。このため、記事中のイベント名候補と検索質問のイベント名が同一のイベントを指していることを判定するため、これを文字列の共通度（共通文字比率）により評価する。

共通文字比率を評価するには、まず抽出したイベント名候補と検索質問のイベント名の双方を形態素解析する。そして形態素単位でその二つを比較し、一致した形態素の文字数の総和を求める。そしてその文字数と検索質問全体の文字数との割合を求め、それが設定した閾値を超えた場合に、同一のイベントを表現していると判断する。現在その閾値は 0.5 に設定している。

4.1.2 「実際にイベントを体験しなければ記述できない表現」の有無による判定

Blog 記事の書き手が実際にイベントを体験した上で記述している場合でも、前節の「その場での体験を示す表現」が明示的に記述されているとは限らない。これは、その場に行ったという事実に言及しなくても、イベントに参加した体験を記述すれば、その場にいたことが明らかになるからである。

このような場合、「その場での体験を示す表現」ではなく、「実際にイベントを体験しなければ記述できない表現」の有無により、実際に書き手がそのイベントを体験しているのかを判断する。この表現は、「楽しかった」「満足だった」等の過去の評価を表す表現*2が多い。

また、「実際にイベントを体験しなければ記述できない表現」が表す書き手の体験は、検索質問の表すイベントで体験したものでなければならない。そこで、検索質問であるイベント名と、blog 記事のタイトルが一致しているという制約を与えている。一致しているかどうかの判断には、前節で述べた文字列の共通度の評価を用いている。以上の流れを図にすると図 3 のようになる。

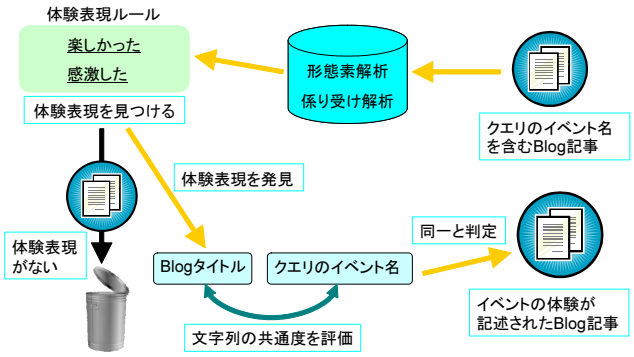

図 3: 「実際にイベントを体験しなければ記述できない表現」の有無の判定

*2 イベントに参加しながら blog 記事を書く場合や、イベントが継続的に参加状態となる性質のものである場合は、評価が過去形にならない。

実際にイベントを体験しなければ記述できない表現

「実際にイベントを体験しなければ記述できない表現」の抽出には、以下のようなルールを用いている。

- 形容詞の連用形（連用タ接続）
  例：よかった、楽しかった、面白かった　等
- 形容詞、形容動詞、副詞、サ変名詞† ＋「でした」「だった」
  例：きれいでした、満足だった　等
- サ変名詞† ＋「する」の連用形＋「た」
  例：感動した、感激した、興奮した、堪能した　等
- 動詞＋接続助詞「て」＋補助動詞「いた」
  例：賑わっていた、盛り上がっていた　等
- 動詞（未然形）＋接尾動詞「れ、られ」＋接続助詞「て」＋補助動詞「いた」
  例：展示されていた、売られていた　等

† 一部の特定のサ変名詞

なお、実際には形態素解析システム茶筌と共に用いるIPADIC[IPA] の品詞体系に基づいた、より詳細なルールとなっている。

4.2 評価

我々の提案するシステムにおける、実体験に基づいて記述された blog 記事かの判定精度を評価した。

評価方法としては、地域ポータルサイトで紹介されていたイベント 10 種について、そのイベント名を Google ブログ検索に投げ、得られた blog 記事の各上位 30 件について、システムに判定させた。

評価結果は、適合率 88.1%、再現率 84.9% であった。なお、Google ブログ検索により得られた記事すべてについて調査した結果、実体験に基づいて記述された blog 記事の割合は 58% であった。

なお、適合率と再現率の定義は以下の通りである。ただし、実体験に基づいて記述された blog 記事のことを実体験記事と呼ぶことにする。

$$\text{適合率} = \frac{\text{出力のうちの実体験記事の数}}{\text{システムが実体験記事とした記事の数}} \quad (1)$$

$$\text{再現率} = \frac{\text{出力のうちの実体験記事の数}}{\text{すべての実体験記事の数}} \quad (2)$$

適合率に比べ、再現率が若干低い値となっている。再現率を上げるためには、評価表現以外の「実際にイベントを体験しなければ記述できない表現」について分析を進める必要がある。

一方で、今後さらに適合率を向上させれば、かなりよい精度で実体験情報をユーザに提示することができるようになる。また、この実体験に基づく記事の言語的な特徴を機械学習させることによって、場が特定されていない記事についても判定ができるようになると考えている。

5. おわりに

Web から実世界を見る手段として、日記の書き手を実世界センサに見立て、日記から実体験に基づく記述のみを抽出する方法について述べた。これは日記を通した実世界センシングであるが、物理デバイスを用いる実世界センシングとはセンシング値の抽象度が異なるため、どのような統合が可能か検討していきたい。

また将来的には、場がクエリからは特定できない場合、例えばクエリが「東京」といった地域名の場合に、有用な実世界情報を提示するための枠組みを別に検討したい。

参考文献

[IPA] 日本語辞書　IPADIC, http://sourceforge.jp/projects/ipadic/

[乾 08] 乾 健太郎, 原 一夫：経験マイニング：Web テキストからの個人の経験の抽出と分類, 言語処理学会 第 14 回 年次大会 (NLP2008) 論文集 C5-4 (2008)

[工藤 02] 工藤 拓, 松本 裕治：チャンキングの段階適用による日本語係り受け解析, 情報処理学会論文誌, Vol. 43, No. 6, pp. 1834–1842 (2002)

[小林 07] 小林 聡, 山田 剛一, 絹川 博之：Blog からのイベント情報の抽出, 第 6 回情報科学技術フォーラム E-038 (2007)

[小林 08] 小林 聡, 山田 剛一, 絹川 博之：Blog からのイベント体験情報の抽出, 電子情報通信学会 2008 年総合大会 D-5-12 (2008)

[倉島 05] 倉島 健, 手塚 太郎, 田中 克己：街 Blog からの体験抽出とその空間的提示手法の提案, 情報処理学会研究報告, Vol. 2005, No. 67 (2005)

[倉島 08] 倉島 健, 藤村 考, 奥田 英範：大規模テキストからの経験マイニング, 電子情報通信学会 第 19 回データ工学ワークショップ (DEWS2008) 論文集 A1-4 (2008)

[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, 電子情報通信学会第 18 回データ工学ワークショップ (DEWS2007) 論文集 A8-1 (2007)

[茶筌] 形態素解析システム 茶筌, http://chasen-legacy.sourceforge.jp/

[南瓜] 係り受け解析システム CaboCha, http://chasen.org/~taku/software/cabocha/