E-038

Blog からのイベント情報の抽出

Extraction of Event Information from Blogs

小林 聡† 山田 剛一‡ 絹川 博之†
Satoru Kobayashi Koichi Yamada Hiroshi Kinukawa

１．はじめに

近年の Blog の普及により、体験した実世界情報を Blog を使って情報発信することが増えている. その発信された実世界情報の中には、実世界で見た対象物や、実世界で何かを体験した際の感想など、実世界で起きている事柄に対する客観的あるいは主観的な情報が書かれている. このような実際に人間の体験を通した実世界情報を得ることは、同じ体験をしようと考えている人達にとって有用であると考えられる.

本研究では実世界情報としてイベントの関連情報を扱う.イベントに参加した人の体験を Blog から抽出することにより、あるイベントに参加したいと思う人が事前にそのイベント参加者の体験を知り、イベントに参加するか否かの判断ができるようになる.

本論文ではこのようなイベント参加者の体験を知り、イベントに参加するか否かの判断ができるようになるシステムを構築するための方法を提案する.

2. Blog エントリに現れるイベント情報の分析

Blog エントリからのイベント情報抽出システムを作るうえで、まず始めに Blog エントリ中にどのようにイベント情報が書かれているかを分析する必要がある.Blog エントリとは Blog を構成する個々の記事のことである.

今回 Blog エントリを分析するにあたりイベント情報を扱っている Web サイトからイベント名を抽出して、Google ブログ検索[1]を用いて、そのイベント名を検索質問として検索した.これによって得られたイベント情報が書かれている Blog エントリ約４００件を分析した.

2.1 イベント関連情報の Blog エントリ

Blog エントリを分析した結果、イベントの関連情報が書かれている Blog エントリは大きく二つのタイプに分類できることが分かった.それを以下に示す

- 実際にイベントへ行き、その感想や見たものなどを記入してあるエントリ
- イベントの紹介や宣伝などが記入してあるエントリ

実際に観察された実世界情報を Blog から抽出するという本研究の趣旨として、今回は「実際にイベントへ行き、その感想や見たものなどを記入してあるエントリ」を研究の対象とする.

† 東京電機大学大学院情報メディア専攻
‡ 東京電機大学/JST-CREST

そして以下の条件が書かれている Blog エントリを実際にイベントに行き、イベントの関連情報が書かれている Blog エントリとする.

（１）Blog 発信者がそのイベントに行ったことが明確に分かる文が書かれている.
例：「（イベント名）に行ってきた」、「（イベント名）を見てきた」等

（２）イベントに行った「感想」や「見たもの」などの表現が書かれている.
例：「楽しかった」、「～が展示されていた」等

これらの条件を満たす Blog エントリを本研究の対象とする.一方、Blog 発信者が実際にイベントに行ったかが分からない Blog エントリや、「感想」や「見たもの」などのイベント関連情報が書かれていない Blog エントリはそのエントリに有益な情報が入っていないと考え今回は対象としない.

例：「（イベント名）が開催されている」、「（イベント名）が始まった」等

2.2 イベント関連情報の項目別分類

実際にイベントに行き、そのイベントの情報が書かれている Blog エントリを収集し、分析することによって、Blog 発信者がイベント情報を書く内容に共通点があることが分かった.以下にそれを項目別に示す.

（１）イベント名

イベント名は主に Blog エントリの始めの文に書かれてある.イベント名の表現には、正式名称がかかれていることもあれば、省略して書かれていることもある.

例：
正式なイベント名：「U2 VERTIGO 2006 TOUR」
省略したイベント名：「U2 のライブ」

（２）見たもの

イベントへ行き、そこに展示されているなど、実際に存在したものが書かれている.イベントに行った際の「見たもの」が書かれているとそのイベントに何があるかといった情報が分かる.

（３）感想

イベントへ行った際に感じた主観的な感想が書かれている. イベントに行った際の「感想」が書かれてあるとそのイベントの評価が分かる.

（４）日付

イベントに行った日付が書かれている.イベントに行った際の「日付」が書かれていることにより、イベントの開催日時が分かり、より詳細な情報が得られる.

「イベント名」はほぼ全ての Blog エントリに書かれている.次に「見たもの」と「感想」については両方書かれ

てある場合もあれば、どちらか一つだけが書かれている場合もある.

これらの項目が書かれている Blog エントリは、そのイベントに関して知りたいと思っている人々に対して、有益な情報が書かれているといえる.そしてこの後解説する提案システムで、ユーザにこの項目を抽出したものを表示すれば、イベントに対して知りたいと思っている人々に対して有益なシステムになると考える.

3. Blog からのイベント情報抽出システム

図１に提案システムの全体の構成を示す.提案システムではまず始めに「Blog エントリの収集」を行う.そして、Blog エントリから「イベントの関連情報の抽出」を行う.

そして最後に抽出した結果を「ユーザに表示」する.

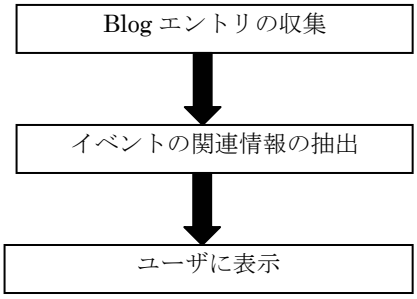

図１．提案システムの構成

3.1 Blog エントリの収集

Blog エントリはエントリ更新情報が集約される ping サーバ、各 Blog サービスのトップページ、既知の Blog を定期巡回し取得する[2].

3.2 イベントの関連情報の抽出

イベントの関連情報を抽出するにあたって、まず始めに Blog エントリの一文ごとに形態素解析と係り受け解析を行う.形態素解析とは形態素の基本形，品詞，接続形等を判別するもので、係り受け解析とは文の係り受け関係を判別するものである.本研究では形態素解析に茶筌[3]、係り受け解析に CaboCha [4]を使用した.

そして、Blog エントリの一文ごとに形態素解析と係り受け解析を行った結果を、GDA（Global Document Annotation）[5]に基づき XML 形式の表現とする.その例を図２に示す.

抽出すべき各項目「イベント名」「見たもの」「感想」「日付」ごとに抽出パターンを作成し、それを基に DOM を使ってマッチングを行い、Blog エントリからイベントの関連情報の抽出を行う.

現在はイベントの関連情報抽出システムを作る上での基礎段階であるイベント名の抽出を行っている.

Blog エントリから「（イベント名）に行ってきた」、「（イベント名）を見てきた」などの文を見つけ、「に」や「を」などの格助詞を持つ格要素をイベント名として抽出する.

具体的には、Blog エントリの一文ごとに形態素解析を行い、文に「行く」「見る」等の連用形（「行ってきた」「見た」等の表現）があるかを判別する.次にその品詞に対応した格助詞（「に」、「を」等）があるかを判別し、存在した場合にはその格要素をイベント名候補として抽出する.そして格助詞にかかっている語句に作成したパターンをマッチングする.これによりイベント名を抽出する.

```
<?xml version="1.0" encoding="Shift_JIS" ?>
- <gda lang="jpn">
 - <su syn="f">
  - <vp syn="f">
   - <adp>
      <ad mph="chasen;助詞+格助詞+連語;;て;テ">て</ad>
     </adp>
     <n mph="chasen;名詞+非自立+一般;;こと;コト">こと</n>
     <v mph="chasen;助動詞;特殊・ダ+連用形;だ;デ">で</v>
     、
    </vp>
  - <adp syn="f">
   - <seg syn="f">
      <seg mph="chasen;未知語;;;">"</seg>
      <n mph="chasen;名詞+一般;;大江戸;オオエド">大江戸</n>
      <n mph="chasen;名詞+一般;;骨董;コットウ">骨董</n>
      <n mph="chasen;名詞+一般;;市;シ">市</n>
      <seg mph="chasen;未知語;;;">"</seg>
     </seg>
     <ad mph="chasen;助詞+格助詞+一般;;に;ニ">に</ad>
    </adp>
    <v mph="chasen;動詞+自立;一段+連用形;出かける;デカケ">出かけ</v>
    <ad mph="chasen;助詞+接続助詞;;て;テ">て</ad>
    <v mph="chasen;動詞+非自立;一段+連用形;みる;ミ">み</v>
    <v mph="chasen;助動詞;特殊・タ+基本形;た;タ">た</v>
    。
  </su>
```

図２．GDA に基づいた XML 表現

3.3 表示インタフェース

現段階ではまだ表示インタフェースに関しては検討段階である.抽出項目である「イベント名」「見たもの」「感想」「日付」をユーザに見やすい形で提示するほか、それらをキーとした検索インタフェースを設計する予定である.

4. おわりに

本論文では Blog からのイベント情報の抽出手法の提案を行った．現在はイベントの関連情報抽出システムを作る上での基礎段階であるイベント名の抽出を行っている.

今後はイベントの関連情報を項目別に分析し、パターンを作成する.そして Blog エントリから項目別にイベントの関連情報を抽出する.さらに検索システムを構築し、実験評価を行う.

参考文献

[1] Google ブログ検索
http://blogsearch.google.co.jp

[2] 小原恭介，山田剛一，絹川博之，中川裕志：Blogger の嗜好を利用した協調フィルタリングによる Web 推薦システム, 第 19 回人工知能学会全国大会, 2C2-02C, 北九州，2005.

[3] 形態素解析システム 茶筌
http://chasen.naist.jp/hiki/ChaSen/

[4] 日本語係り受け解析器 CaboCha
http://chasen.org/~taku/software/cabocha/

[5] GDA (Global Document Annotation)
http://www.i-content.org/gda/