イベントマイニング：出来事を発掘するマイニング手法

数原良彦† 戸田浩之‡ 櫻井彰人†
†慶應義塾大学大学院理工学研究科
‡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
†{suhara, sakurai}@ae.keio.ac.jp ‡toda.hiroyuki@lab.ntt.co.jp

Abstract 本稿では，大量のテキストデータから簡潔な出来事の表現を抽出するイベントマイニングを提唱し，イベントマイニングを実現するための課題を示す．また，現在取り組んでいる研究について概要と手法について述べ，イベントマイニングの今後について論じる．

1. イベントマイニングとは

近年のブログの普及に伴い，ブログを対象にしたサービスの必要性が高まっている．最新の情報がすぐに更新される，ブロガー自身の興味を反映するという理由から，ブログ記事には有益な情報が多く含まれていると考えられ，検索だけではなく，情報抽出への取り組みもなされている．その中のひとつに，ブロガー間で話題になっている語を話題語として抽出する取り組みがある[1]．Technorati は，ユーザたちに興味のあるキーワードを投票させることで話題語を提示している．また，BLOGRANGER や Kizasi.jp では，直近のブログ記事を解析することで話題語を自動的に抽出し，ユーザに提示するサービスを提供している．

しかし，ひとつのキーワードだけでは，話題になっている現実世界の出来事（イベント）についての十分な情報が得られない．ここでイベントとは，文書中で表現されている実世界の出来事や動作のことを示す．BLOGRANGER，Kizasi.jp では，共起頻度の高いフレーズや固有名詞を関連のあるものとしてグループ化し，提示している．これをもとにユーザがイベントを推定することができる．それでも，ユーザが事前知識を持たない限り，提示されたキーワード群からだけではイベントを推測することは難しい．例えば，「イスラエル」「ヒズボラ」「レバノン」が関連する話題語として提示されても，事前知識のないユーザは，現実で起こっているイベントについても内容を知ることができない．

そこで我々は，関連するキーワード間の関係を抽出，提示し，より的確にイベントを表現することで，ユーザの理解を助けることができるのではないかと考えた．ここで言うキーワード間の関係とは，動作，所属，役割，位置，社会的関係などである．図 1 に示すように，大量のテキストデータの中からイベントの内容を簡潔な表現として抽出することをイベントマイニングと呼ぶ．我々の研究では一貫してブログ記事を対象にイベントマイニングの研究を行っているため，ブログ記事からのイベントマイニングのことを狭義のイベントマイニングとして用いる．本稿では，イベントマイニングを実現するために現在取り組んでいる課題と，今後取り組むべき課題について論じる．

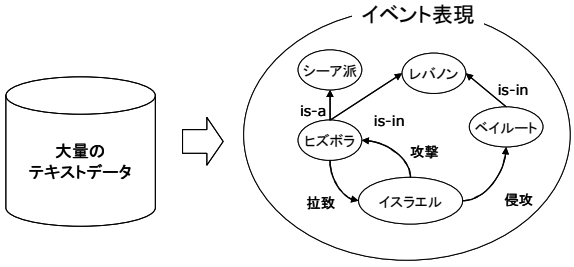

図 1. イベントマイニングの例

大量データの中から知見を得るという目的でデータマイニングという言葉が使われ始めてから 10 年余りが経過した．近年，テキストマイニング，ウェブマイニングをはじめ，さまざまなマイニング手法が提唱されている．従来のデータマイニング，テキストマイニングでは，抽出された解析結果を解釈するために専門家の知識が必要となる場合が多い．

我々の提唱するイベントマイニングでは，専門知識を持たないユーザにも抽出された情報が理解可能で，知見が得られる表現を抽出することを目的としている．

2. 研究のアプローチ

イベントマイニングを実現するため，現在，我々は以下の課題に取り組んでいる．

1. 関連のあるエンティティの抽出
2. エンティティ間の関係の抽出
3. 動作の時系列関係の抽出

2.1 関連のあるエンティティの抽出

イベントマイニングを行うための最初のステップとして，関連のあるエンティティの抽出を行う．図 1 を例にすると，[2]などの既存手法により，ブログ記事で「イスラエル」という頻出キーワードを取得した場合に，「ヒズボラ」や「ベイルート」などの，イベントを表現する上で適切なエンティティを取得する必要がある．

そこで我々は，目的のキーワードと同じ述語に係る格要素は，キーワードと関連が強いのではないかと考え，係り受け関係の解析による関連語の抽出を提案した[3]．

しかし，日本語では前の文脈に現れた語が省略される格要素省略（ゼロ代名詞）と呼ばれる現象が起こるため，必ずしも全ての格要素が取得できるとは限らない．大量のデータが与えられた際には，一文を正しく解析するアプローチだけではなく，簡単な解析を大量の文に対して行い，頻度情報などを利用して解析の精度を高めるアプローチも考えられる．

提案手法では，ブログ記事から選択された大量の文を対象に係り受け解析を行い，＜格要素，格助詞，述語＞の組を抽出する．これを述語パターンと呼ぶ（図 2）．これらの抽出された組の中から，同じ述語をもつものを重ね合わせることで補完を行い，より多くの関連語の抽出を実現する．

図 2. 述語パターンの抽出

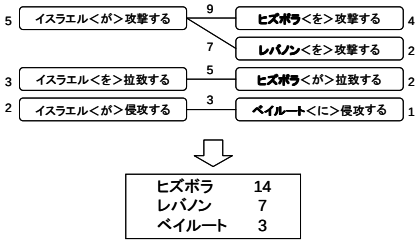

図 3. 述語パターンの組み合わせ

評価実験により，共起する単語を関連のある語として抽出する手法に比べて，提案手法でより多くの関連語を抽出できることが示された[3]．

2.2 エンティティ間の関係の抽出

抽出されたエンティティについて，イベントマイニングの次のステップでは関係の抽出を行う．エンティティ間の関係には，所属，役割，位置，外交関係など様々なものが考えられるが，我々はまず，エンティティ間の動作を表す動作関係に注目し，動作関係をブログ記事から抽出する研究を行った[4]．ここで動作関係とは，動作主，動作対象というふたつのエンティティ間に生じた動作を表す．

我々の提案手法では，関連のあるふたつのエンティティが与えられた際に，それらのキーワードをクエリとする AND 検索を行い，取得したブログ記事の中から動作関係の抽出を行う．

2.1 と同様，文の係り受け関係に注目し，ガ格を動作主，ニ格，ヲ格の格要素を動作対象と見なす．本課題でも格要素省略が問題となるため，2.1 と同様のアプローチで解決を行う．本課題の場合，表層格の意味役割を考慮しているため，同じ述語を持つ述語パターンでも同一の格助詞，ニ格とヲ格の述語パターンの組み合わせは行わない（図 5）．

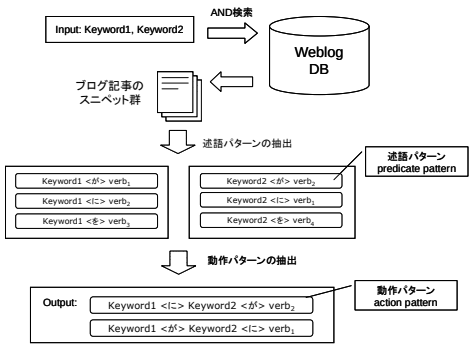

図 4. 動作関係抽出の処理の流れ

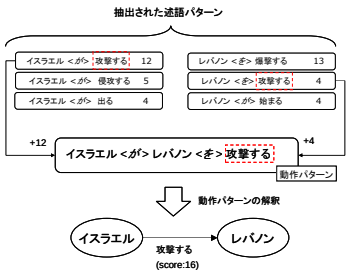

図 5. 動作関係の抽出

評価実験の結果，ひとつの文の中に与えられたふたつのキーワードが格要素として共起するものを解析する手法に比べ，より高い精度で，より多くの適切な動作関係を抽出できることが示された[4].

2.3 関係の時系列情報の抽出

2.2 までの処理では，抽出された関係の時系列情報を考慮していない．図 1 の例では，イスラエルとヒズボラの間の動作関係である「攻撃」「拉致」について，どちらが先に起こったかわからない．どちらの動作が先に起こったかという時間情報を得ることができれば，関係の遷移の可視化，動作の因果関係などを知ることが出来る．このように動作関係の時系列情報を抽出することはイベントの本質的な部分を把握する意味で非常に有益であると考えている．

時系列情報の抽出については近年特に多くの研究が行われている．国内でも動向情報の要約と可視化に関するワークショップが開かれており[5]，時系列マイニングや動向情報の可視化など，大量データの中から動向情報を抽出する既存研究は数多く存在している[6].

時間情報を抽出する方法として，ある段落に含まれるテキストを解析することにより，動作の時間関係を抽出することができると考えられる．例えばスポーツニュースの場合，ある試合の内容について基本的に時系列に沿って記述されるため，動作記述を抽出することができれば，その順番に動作が行われたということが推測できる．しかし，訴訟問題のように長い期間に渡る話題の場合，同じ段落の中で過去の出来事について言及される可能性が高い．また倒置や接続詞によって，必ずしも文の順に動作記述や，それに付随するエンティティが現れるわけではない．これらの問題は自然言語処理の立場からいえば深い解析を行って解決すべき課題であるが，我々はこれまでのアプローチと同様，浅い解析を大量のデータに対して適用することで，この問題の解決を試みる．

具体的には，我々が既存研究で利用した述語パターンを用いた時系列情報の抽出を提案する．テキストの順方向に時間が流れていくと仮定して，述語パターンの絶対的な時間情報（タイムスタンプ）と，述語パターン同士の相対的な出現関係を解析し，動作関係の時系列関係の抽出を試みる．

このように，関係同士の時系列情報を抽出できれば，エンティティ間の関係の時間による変化を表現することが可能となり，ある一時点における静止画のような情報抽出ではなく，連続的な意味を持つ，動画のような情報抽出が可能になると考えている．

本課題については，モデルの検討中の段階で，評価については随時行っていく予定である．

3. イベントマイニングの今後

2 節では現在行っている課題について論じた．本節では今後の課題について概要を述べる．まず，関連のあるエンティティの抽出の発展課題として，ふたつのエンティティの共通のエンティティを発見するという課題が挙げられる（図 6）．図 6 の場合，Person A と Person B の間に何らかの関係が記述されているが，これらの関係が存在せずとも抽出することが望ましい．一見何の関係もないふたつのエンティティが，共通の知人を介してつながりを持っていたということを知ることができ，それだけで非常に興味深い情報となる．

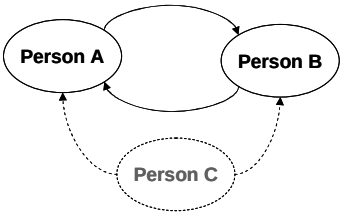

図 6. 第三者の発見

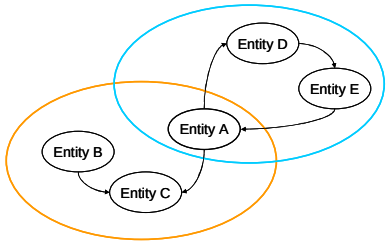

図 7. 複数のイベント表現の判別

また 2 節で紹介した提案手法では，ひとつのキーワードが異なる複数の話題で出現する場合，それらを複数のイベント表現に判別することができない．図 7 に示すように，Entity A が複数の話題で出現する場合，提案手法では，ふたつの異なるイベントがひとつのものとして抽出されてしまう．これも今後の課題である．

4. おわりに

本稿では，大量データの中から出来事の簡潔な表現を抽出することをイベントマイニングと名づけ，イベントマイニングを実現するために必要な課題について論じた．

動作関係の時系列情報を抽出することにより，静止画ではなく動画のように時間的変化のある情報を提示することが可能となる．これにより人間が発掘することの出来ない有用な情報の抽出が可能になると考えている．今後は，関係の時系列情報の抽出を中心に，イベントマイニングの各課題の研究を行う予定である．

参考文献

[1] 関口裕一郎, 川島晴美, 奥田英範, 奥雅博, “ブログ発信者の特徴を利用した話題抽出手法”, DBSJ Letters, vol.5, no.1, pp.9-12, 2006.

[2] H. Isozaki, and H. Kazawa. Efficient support vector classifiers for named entity recognition. In *Proceedings of the 19th international conference on Computational linguistics*, pp.1-7, 2002.

[3] 数原良彦, 戸田浩之, 櫻井彰人, “ブログにおけるイベントマイニングのための適切なキーワード抽出”, 電子情報通信学会第 18 回データ工学ワークショップ, 2007.

[4] Yoshihiko Suhara, Hiroyuki Toda and Akito Sakurai. Event mining from the Blogosphere using topic words. In *Proceedings of the International Conference on Weblogs and Social Media* (*ICWSM* 2007), 2007.

[5] 加藤恒昭, 松下光範, 平尾努, “動向情報の要約と可視化に関するワークショップの提案”, 情報処理学会自然言語処理研究会, 2004-NL-164 (15), pp.89-94, 2004.

[6] 加藤恒昭, 松下光範, 神門典子, “動向情報の要約と可視化-言葉と図で情報をまとめる-”, 情報処理, vol.47, no.9, pp.1013—1020, 2006.