新聞記事と Web からのイベント情報の自動抽出

斉藤隆太 [1,a)]　石野亜耶 [1]　難波英嗣 [1]　竹澤寿幸 [1]

概要：本研究では，新聞記事集合から特定のイベントについて記載された記事を自動判定し，そこからイベント名，開催日時などの情報を抽出する手法を提案する．また，この手法を用いて新聞記事コーパスを解析し，イベントの開催施設名を抽出した後，その場所で開催されるイベント情報が記載された Web ページを自動判定する．提案手法の有効性を確認するため実験を行った結果，イベント記事の自動判定では精度 79.3%が，イベント項目情報の抽出では精度 91.5%が，それぞれ得られた．またイベントページの自動判定では，精度 82.4%が得られた．

Automatic Extraction of Event Information from Newspaper Articles and Web Documents

RYUTA SAITO[1,a)]　AYA ISHINO[1]
HIDETSUGU NANBA[1]　TOSHIYUKI TAKEZAWA[1]

Abstract: In this paper, we propose a method for extracting event information, such as an event name or a schedule from automatically identified newspaper articles, in which particular events are mentioned. We analyze news corpora using our method, and extract facility names from them. Then, we extract Web pages of event schedules of the facilities. To confirm the effectiveness of our method, we conducted several experiments. From the experimental results, we obtained precision of 79.3% in automatic identification of event articles, precision of 91.5% in automatic extraction of event information, and precision of 90.8% in automatic identification of event pages.

1. はじめに

観光は地域における消費増加や雇用の創出など幅広い経済効果をもたらすことから，観光立国の実現を目標に掲げ，2008年10月に観光庁が設置された．その後，観光を21世紀における日本の重要な政策として位置づけた多様な取り組みが推進されている．観光を支援する媒体としては，旅行会社などが運営する観光ポータルサイトや，旅行情報雑誌「るるぶ」などの観光情報データベースなどが作成・公開されている．観光情報の中でも祭り，展覧会，コンサートなどのイベントに関する情報は観光客が行動する際の目的となる重要な情報である．さらに，イベントは日々開催されており，次々に新しいイベント情報が発信される．そのため，観光客などに有益な情報を提供するためには，随時追加されるイベント情報を新たに収集し，データベースを更新することが必要不可欠である．しかし，既存のデータベースでは，人手でイベント情報を抽出し，整理，保守を行うため非常に時間とコストがかかるという問題がある．

そこで本研究では，低コストでのイベント情報データベースを生成し，観光客などに有益な情報提供を行うためのイベント情報検索システムの構築を目標とする．我々はイベント情報の情報源として新聞記事と Web に注目し，新聞記事と Web から手掛かり語を用いた機械学習により，イベント情報の自動抽出を行う．まず，新聞記事について，新聞記事からイベントに関する記事（イベント記事）の判定を行い，判定したイベント記事から，イベント名，開催日時，開催地，開催施設名の 4 種類のイベント項目情報を抽出する．また，新聞記事の見出しからイベント名を含む見出しの判定を行う手法を提案する．

Web について，Web 上でイベント情報の記載されているものとしては，Web ニュースや美術館などの施設が公開しているイベント特設ページが挙げられる．Web ニュースからのイベント項目情報の抽出については，新聞記事のモデルが利用できると考える．そこで，提案手法で作成した新

1 広島市立大学大学院情報科学研究科
Graduate School of Information Sciences, Hiroshima City University
a) saito@ls.info.hiroshima-cu.ac.jp

図 1　イベント情報検索システム

聞記事のモデルを用い，Web ニュースからイベント項目情報の抽出を行う．次に，イベント特設ページからイベント情報の抽出を行うため，本研究ではその第一段階として，イベント情報の記載されている Web ページ（イベントページ）を判定する手法を提案する．本研究では，イベントページを判定するために，過去のイベント記事から抽出した開催施設名を用いる．

本論文の構成は以下のとおりである．2 節では，本研究で提案する手法の成果を利用し，構築したイベント情報検索システムを示し，その動作例を説明する．3 節では関連研究，4 節では提案手法である，新聞記事からのイベント情報の抽出とイベントページの判定について述べる．5 節では提案手法の有効性を調べるために行った実験について述べ，6 節ではそれぞれの実験と結果についての考察を行う．また，結論と今後の課題については 7 節で述べる．

2. システム動作例

本節では，イベント情報を提示するイベント情報検索システムについて，その動作例を紹介する．図 1 は，イベント情報検索システムの画面である．以下では，ある地域におけるイベント情報を検索する場合の一般的な操作手順について説明する．まず，画面上部の検索窓（図中①）に，検索したいイベント名や場所を入力する．（図 1 の場合，場所に“埼玉”という検索語が入力されている）．この状態で検索窓の横にある「search」ボタン（図中②）クリックすると，イベント検索結果の表（図中④）が表示される．イベント検索結果の表は，本研究の提案手法である新聞記事からのイベント情報抽出で作成したモデル（新聞記事からイベント記事を判定する際に用いるモデルとイベント記事からイベント項目情報を抽出する際に用いるモデル）を用いて，Web ニュースからイベント記事を判定し，そのイベント記事から抽出したイベント情報から構成されている．また，記事へのリンクの欄にある見出しをクリックすることで，イベント情報を抽出したイベント記事の本文の内容を閲覧することができる．しかし，新聞記事のモデルを用い，Web ニュースからイベント情報を抽出した場合，新聞では取り上げられないイベント情報は取得できない．そこで，本研究では，より網羅的なイベント情報を取得するため，開催施設名に関するイベント情報の書かれたページを判定する手法を提案する．この手法により抽出されたイベントページはイベント開催施設等の一覧（図中③）にリンクされ，一覧から開催施設名をクリックすることでイベントページを閲覧することができる．図 1 は，イベント名は指定せず，場所のみ“埼玉”と指定し，埼玉県の施設名とイベント情報が結果として表示されている．

3. 関連研究

本研究では新聞記事と Web からイベント情報を抽出している．本研究とはイベントの定義が異なるが，Sakaki ら[1]は地震や台風をイベントと定義し，Twitter ユーザの投稿からリアルタイムなつぶやきを取得する．そして，その Twitter ユーザの位置情報から地震の発生地の検出を行っている．

本研究と同様に Web ニュースからイベント情報を抽出する研究がある．Web などの様々な媒体には未来を予測する記事が掲載されていることから，未来予測は，人々の普遍的欲求であると考えられる．そこで吉田ら[2]は，Web ニュースから未来情報を抽出し，未来情報年表を自動的に構

築する研究を行っている．金澤ら[3]はオンライン文書中から，将来のイベントに関する情報を推測する研究を行っており，ニュース記事の時系列中に現れる周期的なパターンを分析し,将来のイベントを観測する手法を提案している．

本研究では，ユーザ参加型でない低コストのデータベース生成を目指しているが，機械学習とユーザ知識を用い，イベント情報を構造化する研究がある．ウェブ上のドキュメントを機械可読なものとするためには,情報抽出を行い，構造化されたドキュメントに変換する必要がある．そこで森近ら[4]は，機械学習による抽出とシステム利用者の集合知を組み合わせた情報抽出手法を提案している．

ブログから情報抽出する関連研究がある．安村ら[5]は地理情報システムでイベントを扱うことを目的に，イベントの発生時間と発生場所の抽出を行っている．石野ら[6]は旅行ブログが観光情報を得るための有益な情報源であると考え，地域名と土産物の情報抽出を行っている．岡本ら[7]は，随時追加されるイベント情報を取り入れ，有益な情報提供の支援を行うため，地域イベント情報の抽出を行っている．吉田ら[8]は網羅性の高いイベント情報源の作成を目指し，ブログ記事と Web ページからイベント情報抽出法を提案している．いくつかのイベント名に対し，その前後のパターンを用いてイベント名を抽出している. 本研究では，ブログではなく, 新聞記事と Web からイベント情報の抽出を行う．さらに，過去にイベントの開催された開催施設名を用いて Web からイベントページの検出を行い, Web 上からより網羅的にイベントを検出する手法を提案する．

本研究におけるイベント情報抽出の一事例として，会議の情報を抽出する研究がある．Schneider[9] と Issertial ら[10] は“Call for Paper”を情報源とし，会議の名前や開催日などの情報を抽出している．本研究では新聞記事と Web を情報源とし，イベント情報の抽出を行う．

4. イベント検索システムの構築

本研究では，イベント情報を収集する手法として，まず，(1)新聞記事からのイベント情報の自動抽出を行う．そして新聞社の公開する Web ニュースを情報源とし, 新聞記事から作成したモデルを用い,イベント項目情報の抽出を行う．しかし，新聞記事のモデルを用いるため，新聞社以外の公開する様々な Web ニュースを情報源として用いた場合，新聞では取り上げられないイベント情報は取得できない．

そこで本研究では，新聞記事のモデルを用いた Web ニュースからのイベント項目情報の抽出では取得できないイベント情報を補うため，Web から(2)イベントページの自動判定を行う．(1)新聞記事からのイベント情報の自動抽出の手法については 4.1 節，(2)イベントページの自動判定の手法については 4.2 節で説明を行う．

4.1 新聞記事からのイベント情報の自動抽出

本研究では“祭り”や“コンサート”などの観光客や一般の地域住民の参加する行事や催しをイベントと定義する．さらに，イベントに関する情報を含む新聞記事をイベント記事と定義し，イベント記事の判定を行う．ただし，過去のイベントのみ記載されている記事についてはイベント記事として扱わない．イベント記事を判定する際の情報源としては新聞記事を用いる．そして，判定されたイベント記事を情報源として，イベント名，開催日時，開催地，開催施設名の 4 種類のイベント項目情報を抽出する． 4.1.1 節では，新聞記事からイベント記事を自動判定する手法について， 4.1.2 節では，イベント記事からイベント項目情報を自動抽出する手法について，4.1.3 節では，イベント名が含まれる見出しを自動判定する手法について説明を行う．

4.1.1 イベント記事の自動判定

本節では，新聞記事からイベント記事を判定する手法について説明する．イベント記事には，イベント名，開催日時，開催地や開催施設名といったイベント情報が含まれている．人手によりイベント記事と判定した新聞記事の例を図 2 に示す．なお， HEADLINE タグは見出し，TEXT タグは本文を表している．

<HEADLINE>東京ドーム・ふるさとフェア９３　１月２２日から３日間（社告）</HEADLINE>
<TEXT>　わが国最大規模の観光・物産展「東京ドーム・ふるさとフェア’９３」を、今年も一月二十二日から三日間、東京ドームで開催します。</TEXT>

図 2　イベント記事の例

イベント記事を判定するためには，イベント記事やイベント情報特有の言語表現を用いることが有用であると考える．そこで本研究では，図2の例に示されている，“フェア”や“物産展”, “開催します”などのイベント記事によく含まれる語を手掛かり語として収集する．またイベント記事に含まれにくいと考えられる語，例えば，“会議” や “裁判”などの語を不要語として収集する．手がかり語と不要語の収集は著者自身が人手により行う．イベント記事の判定では，機械学習に手掛かり語と不要語の有無を素性として与えることでイベント記事の判定を行う．

機械学習にはTinySVMを用い，手掛かり語と不要語から39の素性を使用した. 使用した素性は大きく以下の6種類に分類される．

- イベント名に関する語(71語)
- 日程に関する語“日”，“まで”など
- “参加費”などのイベントに関連する語(24語)
- 政治・経済に関する語(35語)

- 事件・裁判に関する語(48語)
- その他“写真の有無”，“文の長さ”など

4.1.2 イベント項目情報の自動抽出

本節では，イベント記事からイベント項目情報を抽出する手法について説明を行う．イベント情報にはイベント名，日時，場所，料金，交通などがあり，観光客は主に娯楽を目的としてイベントに参加する．そこで，本研究では，イベント情報の中でも最低限必要な情報であるイベントの名称，日時，場所に着目し，イベント項目情報として，イベント名，開催日時，開催地，開催施設名を抽出する．イベント項目情報を抽出するため，以下の 4 種類のタグを定義する．

- EVENT：開催イベント名
- DATE：イベントの開催日時
- ADDRESS：イベントの開催地
- LOCATION：イベントの開催施設名

また，これらのタグを図 2 のイベント記事に付与した例を図 3 に示す．

<HEADLINE><**EVENT**>東京ドーム・ふるさとフェア９３**</EVENT><DATE>**１月２２日から３日間</DATE>（社告）</HEADLINE>
<TEXT>わが国最大規模の**<EVENT>**観光・物産展「東京ドーム・ふるさとフェア’９３」**</EVENT>**を、今年も**<DATE>**一月二十二日から三日間**</DATE>**、**<LOCATION>**東京ドーム**</LOCATION>**で開催します。</TEXT>

図 3　イベント記事へのタグ付与の例

イベント項目情報を抽出するためには，抽出対象の情報特有の言語表現を用いることが有用であると考える．そこで本研究では，イベント項目情報抽出において，人手によりそれぞれのタグ内やタグ周辺の語を手掛かり語として収集し，それらの有無を素性とした機械学習を行う．イベント項目情報の抽出では, 機械学習としてCRFを用いた. CRF基本手法は与えられた文に含まれる語を分類するのに用いた．素性とタグは以下のようにCRFに与える．

(1) ターゲットとなる単語から，CRFに与える前後の単語数k
(2) ターゲットとなる単語の前に存在する，ターゲットからの距離がk以内に現れる単語
(3) ターゲットとなる単語の後に存在する，ターゲットからの距離が k 以内に現れる単語

本研究では，予備実験の結果から，k=3と定めた．また，機械学習には以下の13種類の素性を使用した．

- 単語
- 品詞
- E_CLUR：“展”，“ライブ”など，イベント名に関する手掛かり語(40語)
- E_CLUE2：EVENTの前後に出現しやすい語
- D_CLUE：“日程”，“日時”など開催日時に関する手掛かり語(63語)
- D_CLUE2：D_CLUEの前後に出現しやすい語
- A_CLUE：“県”，“市”など開催地に関する手掛かり語(97語)
- L_CLUE：“ステージ”，“館”など開催施設名の手掛かり語(72語)
- A_L_CLUE：A_CLUEとL_CLUEの前後に出現しやすい語
- NUMBER：全角数字，漢数字
- KAISAI：“開催”,“行う”などの語
- KAKKO：括弧“「」”，“〈〉”
- KAKKO2：括弧“（）”

4.1.3 イベント名が含まれる見出しの自動判定

本節では，イベント名が含まれる見出しを判定する手法について説明する．イベント記事の見出しには，本文で紹介されるイベントの概要が記載されているものがある．4.1.2節のイベント項目情報の自動抽出でイベント名などが抽出できなかった場合，イベント名が含まれる見出しを表示することで，イベント情報を補うことができる．

イベント名が含まれる見出しを判定するためには，イベント名を含む見出し特有の言語表現を用いることが有用であると考える．そこで本研究では，4.1.1節で人手により収集した手掛かり語の中から，イベント名を含む見出しによく含まれる語を用い，機械学習に手掛かり語の有無を素性として与えることでイベント名を含む見出しの判定を行う．機械学習にはTinySVMを用い，手掛かり語から以下の7種類の素性を使用した．

- “展”
- カギ括弧“「」”
- “開催”
- “フェア”，“コンテスト”などイベント名に関する語
- 文字列＋カギ括弧（“祭り」”，“大会」”など）
- “催し”，“イベント”
- “〇日から”，“来月〇日”など日付に関する語

4.2 イベントページの自動判定

本節では，新聞記事のモデルによる Web ニュースからのイベント項目情報の抽出では取得できないイベント情報を補うため，Web からイベントページの判定について説明を行う．本研究では，観光客や一般の地域住民の参加する行事や催しが記載されている Web ページをイベントページと定義する．ただし，各種観光ポータルサイト内のページと過去のイベントのみが記載されている特設ページについては，本研究ではイベントページとして扱わないこととする．イベントページの例を図 4 に示す．

図 4　イベントページの例

イベントページを判定するにあたり，過去にイベントが開催された開催施設名に関する Web ページには，イベント情報を含むイベント特設ページなどが存在すると考える．そこで本研究では，過去の新聞記事からイベント項目情報の抽出を行い，抽出された開催施設名を用いて，イベントページの判定を行う．しかし，開催施設名の情報だけではイベントページを判定するのは困難であるため，さらに，イベントページによく含まれるキーワード“イベント”を用いる．開催施設名とキーワード“イベント”をクエリとし，Web 検索を行い，検索結果の上位 5 件の Web ページの URL とその Web ページのソースを収集する．検索結果の収集には Yahoo!検索 WebAPI [2] と UNIX コマンドラインツール wget を使用する．

イベントページを判定するためには，イベントページの URLやイベントページ特有の言語表現を用いることが有用であると考える．そこで本研究では，収集したWebページのURLとそのWebページのソースから，イベントページによく含まれる語を手掛かり語として収集する．また，“access”や“youtube”などのイベントページに含まれにくい語を不要語として収集する．手がかり語と不要語の収集は著者自身が人手により行う．イベントページの自動判定では，機械学習に手掛かり語と不要語の有無を素性として与えることでイベントページの判定を行う．

機械学習には TinySVM を用い，手掛かり語と不要語から 7 種類の素性を使用した．(i)～(v)は手掛かり語を含む素性，(vi)は不要語を含む素性である．

(i)　event，calendar，kankoh などの語
(ii)　イベント情報，イベントカレンダーなどの語
(iii)　開催日時に関する語
(iv)　参加費に関する語
(v)　公財，協賛などの語
(vi)　access，youtube，facebook などの語
(vii)　表の有無

5.　実験と結果

本研究では，イベント情報を収集する手法として，以下の2種類の実験を行った．(1)の実験については5.1節，(2)の実験については5.2節で説明を行う．

(1)　新聞記事からのイベント情報の自動抽出
(2)　イベントページの自動判定

5.1 新聞記事からのイベント情報の自動抽出

イベント情報を抽出する際の情報源として，イベント関連の情報を含む記事であるイベント記事を使用する．5.1.1 節では，新聞記事からイベント記事を自動判定する手法について，5.1.2 節では，イベント記事からイベント項目情報を自動抽出する手法について，5.1.3 節では，イベント記事の見出しからイベント名が含まれる見出しを自動判定する手法について実験を行う．

5.1.1 イベント記事の自動判定

イベント記事の自動判定には，読売新聞，朝日新聞，毎日新聞，日経新聞の 4 種類の新聞記事を使用した．これらの 4 種類の新聞記事から“開催”という単語が含まれている新聞記事 2052 件を抽出し，人手によりイベント記事かどうかの判定を行った．人手によりイベント記事と判定した記事を正例とし，人手によりイベント記事と判定されなかった記事を負例として機械学習に用いる．機械学習には TinySVM を用い，2 次の多項式カーネルを使用して 4 分割交差検定を行った．また，“開催”を含むすべての新聞記事をイベント記事として判定した場合の結果をベースラインとし，精度，再現率，F 値を用いて評価を行った．イベント記事の判定の実験結果を表 1 に示す．

表 1　イベント記事の自動判定結果

	精度（%）	再現率（%）	F 値（%）
ベースライン	25.0	100.0	40.0
提案手法	79.3	67.7	73.0

2　http://developer.yahoo.co.jp/webapi/search/websearch/v2/websearch.html

5.1.2 イベント項目情報の自動抽出

イベント項目情報の自動抽出では，人手により 4 種類のタグを付与したイベント記事 416 件を使用した．人手により付与したタグの数を表 2 に示す. 機械学習には CRF を用い，4 分割交差検定を行った．また，単語，品詞のみを用いて機械学習を行った結果をベースラインとし，精度，再現率，F 値を用いて評価を行った．イベント項目情報の自動抽出結果を表 3 に示す.

表 2　人手により付与したタグの数

タグの種類	個数（個）
EVENT	693
DATE	657
ADDRESS	658
LOCATION	464

表 3　イベント項目情報の自動抽出結果

	ベースライン（単語，品詞）			提案手法（ベースライン＋手掛かり語）		
	精度（%）	再現率(%)	F 値（%）	精度（%）	再現率(%)	F 値（%）
EVENT	89.9	51.0	65.1	91.2	67.3	77.4
DATE	96.5	79.9	87.4	96.8	85.5	90.8
ADDRESS	91.0	76.9	83.3	90.8	81.6	86.0
LOCATION	89.6	54.7	67.9	87.3	69.3	77.3
平均	91.8	65.6	76.5	91.5	75.9	83.0

5.1.3 イベント名が含まれる見出しの自動判定

イベント名が含まれる見出しの自動判定には，イベント記事の見出し 416 件を人手により判定した結果を機械学習に用いる．機械学習には TinySVM を用い，2 次の多項式カーネルを使用して 4 分割交差検定を行った．また，イベント記事の見出し 416 件のすべてにイベント名が含まれるとして判定した場合の結果をベースラインとし，精度，再現率，F 値を用いて評価を行った．イベント名が含まれる見出しの自動判定の実験結果を表 4 に示す.

表 4　イベント名が含まれる見出しの自動判定結果

	精度（%）	再現率（%）	F 値（%）
ベースライン	65.1	100.0	78.9
提案手法	81.6	86.4	83.9

5.2 イベントページの自動判定

イベントページの自動判定には，“開催施設名　イベント”をクエリとして Web 検索を行い，検索結果の上位 5 件の Web ページの URL とその Web ページのソースを収集した．そして，人手によりイベントページかどうかを判定した Web ページ 1022 件を機械学習に用い，イベントページの判定を行った．機械学習には TinySVM を用い，2 次の多項式カーネルを使用して 4 分割交差検定を行った．また，人手によりイベントページかどうかを判定した Web ページ 1022 件のすべてをイベントページとして判定した場合の結果をベースラインとし，精度，再現率，F 値を用いて評価を行った．イベントページの自動判定の実験結果を表 5 に示す.

表 5　イベントページの自動判定結果

	精度（%）	再現率（%）	F 値（%）
ベースライン	25.8	100.0	41.0
提案手法	82.4	52.2	63.9

6. 考察

6.1 イベント記事の自動判定結果からの考察

本節では，提案手法により誤って判定を行った新聞記事について考察する．まず、精度低下の原因について考察を行う. 誤ってイベント記事と判定された新聞記事の例を図5に示す.

<TOPIC>**催し**案内</TOPIC>
<HEADLINE>短信　／千葉</HEADLINE>
<TEXT>
市川市が懇話会委員を募集　市内在住で、子ども、子育てに関する幅広い知識を持ち、１～２カ月に１回程度昼間に**開催する**懇話会に出席できることが条件。任期は**０５年３月３１日まで**。
＊＊＊＊＊＊＊＊＊＊＊　略　＊＊＊＊＊＊＊＊＊＊＊
書類選考の上、面接試験がある。**問い合わせ**は同課（電話０４７・３３４・１１７７）。
</TEXT>

図 5　提案手法により誤って判定された新聞記事の例

図5において，太字がそれぞれ手掛かり語である．イベント記事の手掛かり語となる“催し”や“開催する”などの語が複数含まれているため, 誤って判定されたと考える．誤って判定された新聞記事の特徴としては，イベントと同様に開催日や開催施設名を記事内に含むものが多い．

次に，再現率低下の原因について考察を行う．提案手法により正しくイベント記事と判定できなかった新聞記事には，イベント名や施設名が固有名詞であるなどの理由で，手掛かり語に含まれていない，または，不要語を含んでいるという特徴があった．よって，イベント記事と判定できる特徴が得られないため，提案手法により判定できなかったと考える．

6.2 イベント項目情報の自動抽出結果からの考察

イベント項目情報の抽出結果について，(1)提案手法により誤って抽出した例と，(2)提案手法により抽出できなかった例について分析を行う．

(1) 提案手法により誤って抽出した例

人手ではタグをつけなかったが，提案手法により誤ってタグを付与した例について分析を行った．分析の結果，精度低下の原因は主に，以下の2種類である．

(i) 過去に開催されたイベントや関連イベントの情報

(ii) 手掛かり語の問題

(i)について，誤って抽出したものとして過去のイベントや関連イベントが挙げられる．本研究では，タグとその周辺の手掛かり語を用いて情報を抽出している．したがって，開催されるイベントと過去のイベントや関連イベントの区別ができず，誤って判定してしまったと考えられる．これらの誤りは，提案手法により抽出されたイベント項目情報を含む文が過去の表現を含むかどうか，または，その周囲に存在する日付と新聞記事の日付を比較することで改善できると考える．

(ii)について，日付や地名は記事中に頻出する語である．本研究ではDATE，ADDRESSの手掛かり語として“日”や“市”などを用いているため，イベントに関する日付や地名以外も判定されてしまったと考える．そのため，システムを構築する際には，判定されたDATEやADDRESSの周囲にEVENTが存在するかどうかを調べ，EVENTが周囲に存在しないDATEなどは取り除くなどの処理を行い誤って判定されたものを取り除く必要があると考える．

(2) 提案手法により抽出できなかった例

提案手法により抽出できなかった事例について分析を行う．結果，原因は主に手掛かり語の素性だけでは抽出が困難であるものであった．誤り例を表6に示す．表において，太字がそれぞれのタグの手掛かり語，下線部分が再現できなかった部分である．

表 6　提案手法により抽出できなかったタグの例

EVENT	歌人、原阿佐緒**展**では、
DATE	「ふるさとフェア」22**日**開幕
ADDRESS	二月一日まで、大丸・梅田の十一階
LOCATION	午後６時３０分開演　神戸文化小**ホール** ＪＲ神戸駅から東へ徒歩

EVENT の例について，手掛かり語は“展”のみである．例のように，イベント名に固有名詞が含まれており，カギ括弧などでイベント名が囲まれていない場合，イベント名と判定できる十分な特徴が得られないため提案手法により抽出できなかったと考える．

DATEの例について，本研究ではDATEとその前後に存在する助詞などを手掛かりとし，抽出を行っている．したがって，抽出対象の前後に助詞が存在せず，日付のみ記載されている場合，開催日時と判定できる特徴が得られず，提案手法により抽出できなかったと考える．

ADDRESSの例について，本研究ではADDRESSの手掛かり語として“県”や“市”などを用いている．しかし，“県”や“市”などを用いて表記しない地名に関しては手掛かり語が存在しないため，提案手法により抽出できなかったと考える．

LOCATIONの例について，“ホール”のみが手掛かり語である．本研究では，イベント記事を形態素解析し素性を与えている．したがって，施設名の前後に空白が含まれている場合などは，タグの前後に手掛かりとなる助詞が存在しない．よって，施設名に含まれる一部の手掛かり語のみでは，施設名と判定できるほどの十分な特徴が得られず，提案手法により抽出できなかったと考える．

6.3 イベント名が含まれる見出しの自動判定結果からの考察

本節では，提案手法により誤って判定を行った見出しについて考察する．誤ってイベント名を含むと判定された見出しは，過去に行われたイベントなどに関する内容の見出しであった．本研究で用いる手掛かり語を含んでいるため，誤って検出されたと考えられる．

次に，再現率低下の原因について考察を行う．提案手法により正しく判定できなかった見出しには，固有名詞を含むイベント名が含まれており，本文の内容を考慮しなければ判断できないものであった．よって，手掛かり語が存在せず正しく判定できなかったと考える．

6.4 イベントページの判定結果からの考察

イベントページの判定結果について，(1)提案手法により誤ってイベントページと判定された例と(2)提案手法により正しくイベントページと判定できなかった例について分析を行う．

(1) 提案手法により誤ってイベントページと判定された例

提案手法により誤ってイベントページと判定された例を図6に示す．

図 6　誤ってイベントページと判定された例

図6は提案手法により誤って判定したWebページのURLとそのページ内の一部分を示している．図6は観光ポータルサイトであり，イベント特設ページが存在する．イベント特設ページのURLは，本研究で定義したイベントページのURLによく含まれる“event”，“calendar”などの手掛かり語を含んでいる場合が多い．さらに，ページ内にも“開催日”，“開催場所”などの手掛かり語を含むため，誤ってイベントページと判定されたと考える．

(2) 提案手法により正しくイベントページと判定できなかった例

提案手法により正しくイベントページと判定できなかった例を図7に示す．

図7　正しくイベントページと判定できなかった例

図7は，提案手法により正しく判定できなかったある施設のトップページの一部分を抜き出したものである．図7のようにWebページの一部分にイベントに関する情報が記載されている場合，そのWebページのURLには“event”や”calendar”などのイベントページのURL特有の手掛かり語は存在しない．また，Webページ内にもイベントに関する情報が少なく簡潔に記載されていることが多い．よって，イベントページと判定できるほどの手掛かり語が存在しないため，提案手法により正しく判定できなかったと考える．

7. おわりに

本研究では，低コストでのイベント情報データベースを作成し，イベント情報検索システムの構築を行うため，手掛かり語を用いて自動的にイベント情報を収集する手法を提案した．まず，新聞記事がイベント記事かを自動判定する手法，次に，イベント記事からイベント項目情報を表すためのタグ“EVENT”，“DATE”，“ADDRESS”，“LOCATION”を定義し，イベント項目情報を自動抽出する手法を提案した．さらに，イベント名が含まれる見出しを自動判定する手法を提案した．そして，イベント情報検索システムを作成するにあたり，より網羅的なイベント情報収集のため，過去の新聞記事からイベントの開催施設名を収集した．その開催施設名を用い，Webからイベントページの判定を行った．提案手法には，それぞれの手法に関する手掛かり語を用いた機械学習を行った．実験の結果，イベント記事の自動判定では精度79.3%，再現率67.7%，F値73.0%，イベント項目情報の自動抽出では，精度91.5%，再現率75.9%，F値83.0%，イベント名が含まれる見出しの自動判定では，精度81.6%，再現率86.4%，F値83.9%という結果が得られた．また，イベントページの自動判定については精度82.4%，再現率52.2%，F値63.9%という結果が得られた．

今後の課題として，イベントページからのイベント項目情報の抽出を行うことによる，イベント情報検索システムの改良が挙げられる．イベントページにはカレンダーや表でイベントを表記しているページが多く存在する．したがって，手掛かり語を用いたイベント項目情報の抽出のほかに，表の解析を行うことでより網羅的な情報抽出を行う．

参考文献

[1] Takeshi Sakaki，Makoto Okazaki，and Yutaka Matsuo: *Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors*，Proc.18th International WorldWide Web Conference(WWW2010), 2010.

[2] 吉田光男，乾孝司，山本幹雄：Webニュースを用いた未来情報年表の自動構築，第3回楽天研究開発シンポジウム，2010.

[3] 金澤健介，Adam Jatowt，小山聡，田中克己：Webからの将来情報の発見･分析にむけて，情報処理学会研究報告，データベース・システム研究会報告2008(88)，pp.325-330，2008.

[4] 森近憲行，濱崎雅弘，亀田尭宙，大向一輝，武田英明：機械学習とユーザ知識を用いたイベント情報の構造化，人工知能学会全国大会(第24回)論文集，No.1D2-3，2010.

[5] 安村祥子，池崎正和，渡邉豊英，牛尼剛聡：blogマッピングを用いたイベント情報抽出，DEWS2007，B7-10，2007.

[6] 石野亜耶，難波英嗣，田熊遥，尾崎貴紘，小林大祐，竹澤寿幸：旅行ブログからの観光情報の自動抽出，知能と情報，Vol.22，No.6，pp.667-679，2010.

[7] 岡本昌之，菊池匡晃：ブログからの地域イベント情報抽出，情報処理，Vol.51，No.1，pp.14-17，2010.

[8] 吉田将人，福原知宏，増田英考：ブログ記事とWebページを用いたイベント情報抽出手法の提案，情報処理学会研究報告，デジタルドキュメント2009(35)，pp.37-44，2009.

[9] Karl-Michael Schneider*: An Evaluation of Layout Features for Information Extraction from Calls for Papers*, LWA2005, pp.111-115, 2005.

[10] Laurent Issertial and Hiroshi Tsuji: *Information Extraction and Ontology Model for a 'Call for Paper' Manager*, iiWAS2011，pp.539-542, 2011.