社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2009－ＦＩ－94 (6)
2009－ＤＤ－70 (6)
2009／3／25

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

吉田 将人 †, 福原 知宏 ††, 増田 英孝 †

† 東京電機大学未来科学部 †† 東京大学人工物工学研究センター

ブログ記事とWebページを利用したイベント情報抽出手法を提案する．提案手法は，ブログ記事からイベント名抽出パターンを構築し，Webページからイベント名を抽出する．本研究では，ブログ記事とWebページを利用したイベント情報抽出手法を提案する．ブログ記事を用いることにより，記事の書かれた日付が分かり，イベント名抽出パターンとイベント開催日の関係を把握できる．Webページを用いることにより，イベント名検索の網羅性を広げることができる．提案手法では，まず，いくつかのイベント名に対してブログ記事を収集し，そこからイベント名の前後に連接しやすいパターンを抽出する．次に，抽出したパターンを用いてWeb全体からイベント名を収集する．提案手法のイベント名収集適合率と将来構想について報告する．

An Extraction Method of Event Names using Blog and Web Pages

Masato YOSHIDA † Tomohiro FUKUHARA †† Hidetaka MASUDA †

† School of Science and Technology for Future Life,Tokyo Denki University

†† RACE (Research into Artifacts, Center for Engineering),University of Tokyo

An extraction method of event names appeared on the Web using blog and Web articles is described. Proposed method extracts event names from Web pages by finding extraction patterns of event names from blog articles. The method finds extraction patterns from blog articles that contain event names given by a user. Because different names for the same event can be appeared on the Web, the method identifies the same event using a string kernel that can measure similarities of event names. Then, the method finds event names by using extracted patterns. Preliminary results of an experiment are described.

1 はじめに

今日，展示会，音楽フェスティバル，美術展など，さまざまなイベントが日々開催されている．しかし，生活の中でイベント情報を知る機会は限られている．例えば，Webサイトや電車の中吊り，雑誌の広告で見掛ける程度である．このように日常生活でイベント情報を知る機会は限られている．また，Web上からイベント情報を得ようとすると，一つのサイトにイベント情報がまとまって掲載されていないため，複数のサイトを閲覧しなければならない．例えば，毎年，明治神宮外苑前で「100% Design Tokyo」という，比較的規模の大きいデザイン系イベントが開催されている．しかし，Google検索[1]で「イベント情報」と検索して得られたイベント情報掲載サイトの上位３件にはこのイベントの情報は載っていない．このイベントの情報を得るには，デザイン系イベントを専門に扱うサイトを閲覧する必要がある．

そこで本研究では，網羅性の高いイベント情報源の作成を目指した，ブログとWebページを用いたイベント情報抽出手法を提案する．イベント情報を一つに集約することで，ユーザがイベント情報を知る機会を増やすことが出来る．提案手法では，いくつかのイベント名に対してブログ記事を収集し，そこからイベント名の前後に連接しやすいパターンを抽出する．ブログ記事を用いることにより，記事の書かれた日付が分かり，イベント名抽出パターンとイベント開催日の関係を把握できる．次に，抽出したパターンを用いてWeb全体

[1] http://www.google.co.jp

からイベント名を収集する．本手法により，主要なイベント情報に加え，上記の情報掲載サイトに掲載されていないイベントも取得できる．提案手法のイベント名収集適合率と将来構想について報告する．

本論文の構成は以下のようになっている．第2節は提案手法について，第3節はブログからのイベント名抽出について，第4節は提案手法の評価について，第5節は実験の評価と今後の展開について，第6節は本研究のまとめについて述べる．

2 網羅性の高いイベント情報抽出手法の提案

現在用いられている網羅性の高いイベント情報源作成アプローチと，提案手法の特徴について述べる．

2.1 網羅性の高いイベント情報源の作成

網羅性の高いイベント情報源の作成には，複数のアプローチがある．その中でも以下の2つの手法に付いて述べる．

(1) 人手によるイベント情報の収集
(2) 複数イベント情報サイトの情報集約

(1) 人手によるイベント情報の収集

現在，ソーシャルイベントサイト「eventcast[2]」という，ユーザが自由にイベント情報を登録し，共有できるサイトがある．人手によるイベント情報収集は高い適合率が期待できる．事実，ユーザに興味のあるイベントであれば、「100% DesignTokyo」というアート系イベントから，ウィンタースポーツグッズの展示販売会「WINTER SPORTS FESTA」までこのサイトには掲載されている．しかし，人手によるイベント情報の収集には，ユーザの興味に偏りがあるため，提案手法で抽出できたイベント「ふくろ祭り[3]」は掲載されていなかった．この手法の欠点は，このサイトを利用しないユーザからの情報は集まらない，という点である．

(2) 複数イベント情報サイトの情報集約

2008年7月,『イベント“発見”サイト「ことさが[4]」』がリリース[5]された．このサイトの特徴は，『独自開発のクローラーで，イベント情報を更新している約700サイトから自動収集し，ユーザーがさまざまなイベントを“発見”できる』という点である．複数のイベント情報サイトに掲載されている情報を集約することで網羅性の向上を図っている．しかし，こちらのサイトにも地域に根付いたイベント「ふくろ祭り」は掲載されていなかった．この手法の欠点は，抽出元のイベント情報掲載サイトに掲載されていないイベント情報は取得することができない，という点である．

2.2 ブログとWebページからのイベント名抽出

以上の結果を踏まえ，本研究ではブログとWebページを用いたイベント情報抽出方法を提案する．今日，ブログサービスは幅広く普及し，誰でも簡単にイベントに対して記事を作成できる．よって，ブログを抽出対象に含む事でイベント名抽出の網羅性を広げることができる．提案手法は，ブログ記事からイベント名抽出パターンを構築し，Webページからイベント名を抽出するという手法である．次節より，提案手法について詳しく述べる．

3 ブログからのイベント名の抽出

この節では，ブログからのイベント名抽出手法について述べる．

3.1 ブログ中に存在するイベント名

ブログ記事にイベント名がどのように記述されるか観察してみると，

- ... 絵画館前広場 で催されている「100% Design Tokyo」に行ってきました...
- ... 東京都美術館で で開催されている,「フェルメール展」に行ってきました...
- ... 幕張メッセ で開かれている「 CEATEC JAPAN 2008 」に行ってきました...

のように，イベント情報の表現にはある程度の共通性が見られる．ところで，例にあげた文中の“開催されている”等のイベントに係る表現はどのくらいの頻度で用いられるのだろうか．ここで,「東京国際アニメフェア2008」というイベント名を含むWebページ(ブログ記事を含む)に対して，イベントを記述する際の表現として想起された“開催されている”，“開催された”，“行われた”，“開催される”がどれぐらい含まれるのかを調べた．Table 1に結果を示す．また，調査時の「東京国際アニメ

[2] http://clip.eventcast.jp/
[3] 「ふくろ祭り」とは毎年9月下旬に東京都池袋で開催される，地域のお祭りである．
[4] http://cotosaga.com
[5] イベント“発見”サイト「ことさが」 - ITmedia News http://www.itmedia.co.jp/news/articles/0807/30/news038.html

Table 1　「東京国際アニメフェア」を含む記事中の表現と件数

イベント記述表現	ヒット件数
開催されている	1,899
開催された	2,517
行われた	1,496
開催される	1,804

フェア 2008」を含む Web ページ数は 27,629 件であった．

このように，イベント名を記述する際の普遍的な表現が存在すれば，それを基に抽出パターンを作りイベント名を抽出することが出来る．イベント名を明記する際の普遍的な表現があるか調べるために，次の調査を行った．

3.2　イベント名の前後に連接する定型的表現の調査

イベント名の前後に連接する定型的表現を調査する．調査により定型的表現にはある程度の共通性があることがわかった．

3.2.1　調査手順

イベント名の前後にはどのような文字列が連接するのかについて以下の手順で調査をした．Fig.1 に調査の流れを示す．

(1) イベント名の収集
(2) イベント名を含む記事の収集
(3) ブログ記事の本文の特定
(4) 文字列カーネルを用いた記事中のイベント名の特定
(5) 記事中に存在するイベント名の前後に連接する文字列の抽出
(6) Kiwi アルゴリズムによる定型的表現の取得

以下，各手順について説明する．

(1) イベント名の収集

イベント名をキーワードとして Google ブログ検索で検索した際に数千件の検索結果が得られたイベントを 21 件選んだ．Table 2 に調査に用いたイベント名と取得記事数を示す．

(2) イベント名を含む記事の収集

Yahoo!ブログ検索の検索結果を利用する．各イベント名をクエリとして検索し，検索結果 RSS よりイベント名毎に記事を 1,000 件まで取得する．

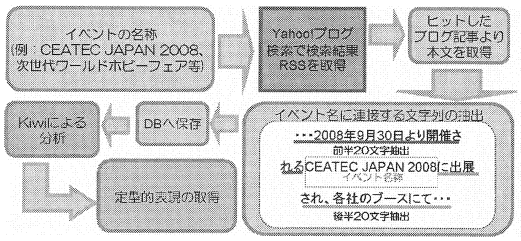

Fig. 1　連接する文字列調査の流れ

(3) ブログ記事の本文の特定

本研究では，1 文中に存在するイベント名の前後に連接する文字列を抽出する．このため，文字列抽出の対象をブログ記事の本文に限定する必要がある．本研究では，RSS から得られる本文要約文を元にブログ記事の本文を特定する．記事の HTML ソースから本文要約文を囲う DIV タグを探し，DIV タグ中の文字列を本文として用いた．

(4) 文字列カーネルを用いた記事中のイベント名の特定

イベント名の表記ゆれを解消するために文字列カーネル (String Kernel) を用いた．Yahoo や Google などの Web 検索エンジンは検索の網羅性を上げるために，空白や特殊記号を無視してクエリがマッチする記事を返す．数字の全角・半角もまた，区別されない．例えば「アフリカンフェア」と検索すると，記事の本文中に

- 「アフリカン・フェア」(記号が存在)
- 「アフリカンフェアー」(表記揺れ)

等，クエリと完全に一致しない文字列を含む記事も結果として得られる．このような場合，記事本文中からイベント名を完全一致で探そうとすると，適切にイベント名の前後文字列を抽出することができない．

本研究では，**文字列カーネル**(付録 (A) 参照) を用いてイベント名と，本文中に存在する助詞や助動詞などで切り出した単語列との文字列類似度を計算し，一定の類似度を持つ単語文字列をイベント名と判断する．また，英数字の全角と半角やひらがな，カタカナなどを区別せずに類似度を計算する．文字列カーネルを適用することで，本文中の表記が異なって書かれているイベント名を発見出

Table 2　調査に用いたイベント名と取得記事数

イベント名	取得記事数 (件)
コミックマーケット	1000
ジャンプフェスタ 2009	140
ジャパンドッグフェスティバル 2008	25
100% design tokyo	468
ワンダーフェスティバル 2008	977
CEATEC JAPAN 2008	996
次世代ワールドホビーフェア	999
アートフェア東京 2008	95
セミコンジャパン 2008	44
東京国際アニメフェア 2008	999
インポートオートサロン 2009	28
フェルメール展	998
Pet 博 2008	141
ルノワール+ルノワール展	378
第 11 回文化庁メディア芸術祭	930
Web2.0 EXPO	540
AOU2008 アミューズメントエキスポ	133
第 40 回東京モーターショー	996
COUNTDOWN JAPAN08/09	394
NHK 杯国際フィギュアスケート競技大会	113
エコプロダクツ 2008	927

来る．今回は文字列カーネルの中でも Lodhi[2] らが提案した，“文字間のギャップを許容する” 文字列カーネルを用いた．

(5)Kiwi アルゴリズムによる定型的表現の調査

ブログ記事から切り出したイベント名に連接する文字列を Kiwi アルゴリズム [1] を用いて解析する (付録 (B) 参照)．Kiwi アルゴリズムとは，田中らが提案した定型的表現抽出アルゴリズムである．イベント名に後続する文字の種類数が増加するところまでを切り出すことにより，定型的表現を取得する．本研究では，イベント名を含む記事群に対して Kiwi アルゴリズムを適用することで，イベントを記述する際の前後の定型的表現を取得する．

3.3　調査結果

今回の調査では，収集したブログ記事より 9,284 件のイベント名に連接する文字列を抽出した．調査の結果例として，「CEATEC JAPAN 2008」と「次世代ワールドホビーフェア」の例を Table 3 と Table 4 に示す．表中の「頻度」は，各定型的表現が，例のイベント名を含む記事より得られたイベント名に連接する文字列群中で用いられていた回数である (「CEATEC JAPAN 2008」で得られた連接する文字列件数は 614 件，「次世代ワールドホビーフェア」では 719 件であった)．

以下にイベント名の前後に続く定型的表現の特徴を示す．

(1) イベント名の定型的表現にはある程度の共通性がある．
(2) イベント名は鉤括弧 (「」) に囲まれて書かれやすい
(3) “開催” という単語がフレーズに多く含まれることが多い．

(1) に関しては，前部が “**開催される「**”，後部が “」” という定型的表現が，例に挙げた「CEATEC JAPAN 2008」と「次世代ワールドホビーフェア」の結果に共通して出現していることからわかる．また，この定型的表現は調査に用いた 21 件のイベント中 10 件のイベントで用いられていた．

(2) に関しては，調査に用いた 21 件のうち，20 件において鉤括弧から始まる定型的表現が一番多く抽出された．

また，(3) に関しては，“開催” を含む表現が用いられるイベントは 21 件中 20 件であった．

このように，イベント名を記述する際の定型的表現が存在し，ある程度の共通性を持つことがわかった．次に，この調査で得られた定型的表現を元にイベント名が抽出できるか実験した．

4　評価実験

定型的表現群からイベント名抽出パターンを選定し，データセットに対して抽出実験を行った．選定したイベント名抽出パターンの適合率と網羅率を算出し，結果について考察する．

4.1　実験手順

以下の手順で実験を行った．

(1) 定型的表現群からのイベント名抽出パターンの選定
(2) データセットの作成
(3) パターンによるデータセットからのイベント名候補文字列の抽出

Table 3 「CEATEC JAPAN 2008」の定型的表現の結果例

前部	EVENTNAME	後部	頻度
「	CEATEC JAPAN 2008	」	119
「	CEATEC JAPAN 2008	」に	31
「	CEATEC JAPAN 2008	」の	27
る「	CEATEC JAPAN 2008	」	25
れる「	CEATEC JAPAN 2008	」	18
「	CEATEC JAPAN 2008	」で	18
開催される「	CEATEC JAPAN 2008	」	17
展示会「	CEATEC JAPAN 2008	」	15
展「	CEATEC JAPAN 2008	」	15
総合展「	CEATEC JAPAN 2008	」	14

Table 4 「次世代ワールドホビーフェア」の定型的表現の結果例

前部	EVENTNAME	後部	頻度
「	次世代ワールドホビーフェア	」	103
る「	次世代ワールドホビーフェア	」	23
「	次世代ワールドホビーフェア	」に	21
「	次世代ワールドホビーフェア	」開	17
た「	次世代ワールドホビーフェア	」	15
れる「	次世代ワールドホビーフェア	」	15
「	次世代ワールドホビーフェア	」開催	14
される「	次世代ワールドホビーフェア	」	13
「	次世代ワールドホビーフェア	」が	13
開催される「	次世代ワールドホビーフェア	」	12

(4) 適合率と網羅率の測定

以下，各手順について説明する．

(1) イベント名抽出パターンの選定

まず，3.3 節の調査から得られた 59,043 件の定型的表現を次の条件で選定する．

(a) 3.3 節の結果より，前後共に鉤括弧 (「」) から始まらない定型的表現を削除する．
(b) (a) で残った定型的表現の内，21 件中 4 件以上のイベントで用いられていない定型的表現を削除する．(以下，ある定型的表現が 21 件中何件のイベントで用いられているか，を**被イベント使用数**と呼ぶ．)
(c) (b) で残った定型的表現の内，定型的表現の文字列数が前後共に鉤括弧 (「」) を含み 3 文字以下の定型的表現を削除する．
(d) (c) で残った定型的表現の内，包含関係にあるような定型的表現は，被イベント使用数が互いに同値である場合は文字長が長い定型的表現へ統合し，同値でない定型的表現は被イベント使用数が高い定型的表現へ統合する．
(e) (d) で残った定型的表現の内，3.3 節の結果より，“開催” を含まない定型的表現を削除する．
(f) (e) で残った定型的表現の内，21 件のイベントで得られた 9,284 個のイベント名に連接する文字列群における出現頻度が 10 未満である定型的表現を削除する．

こうして，8 件のイベント名の定型的表現を選定した．ここで，この 8 件をイベント名抽出パターンと呼ぶ．また，ベースラインとして，鉤括弧のみの抽出パターン (以下，**ベースライン抽出パターン**) も対象に加えて実験を行った．

(2) データセットの作成

実験に用いるデータセットを用意する．まず，3.3 節で用いたイベント 21 件以外のイベント (以下，対象イベント名) を 29 件集め，各イベントに対し記事を Yahoo!ブログ検索[6]より 50 件ずつ取得する．また，イベントでないキーワードを同じく 29 個集め，各キーワードに対し記事を同じく Yahoo!ブログ検索より 50 件ずつ取得する．合計 2,900 記事をデータセットとした．

(3) イベント名候補文字列の抽出

パターンの前部文字列と後部文字列をマッチング文字列に指定し，最短マッチでパターンの間に出現する文字列を抽出する．抽出した文字列中に読点が存在した場合は不適合と判断した．

(4) 適合率と網羅率の測定

本研究ではイベント名抽出の評価に適合率と網羅率の尺度を用いる．式 (1)，(2) にそれぞれ適合率，網羅率を示す．

$$\text{網羅率} = \frac{\text{対象イベント名を抽出できた記事数}}{\text{対象イベント名を含む記事数}} \times 100[\%] \quad (1)$$

[6] http://blog-search.yahoo.co.jp/

$$適合率 = \frac{対象イベント名を抽出できた記事数}{抽出されたイベント名候補数} \times 100[\%] \quad (2)$$

正確な網羅率を測定するには，ブログ空間に存在するイベント名を含んだブログ記事と含んでいないブログ記事の割合と等しいブログ記事集合を用意する必要がある．しかし，今回は，対象イベント名を含むブログ記事から，抽出パターンを用いて対象イベント名を抽出できた記事の割合を網羅率とする．

また，適合率は，ある対象イベント名を含む記事よりその対象イベント名が抽出できた記事を正解と判断する．抽出された文字列が，実在するイベント名であるが，抽出された記事が含む対象イベント名で無い場合は無視し，抽出された文字列が確実にイベント名でない場合を不正解と判断する．よって，適合率の計算式の分母は，抽出できた文字列数から対象イベント名でないイベント名が抽出できた記事数を引いた値になる．今回はこれを「抽出されたイベント名候補数」とし，式中に用いる．

4.2 実験結果

Table 6 に各イベント名抽出パターンの適合率とイベント名候補文字列抽出数を示す．実験の結果，網羅率は 3.2%，適合率は 97.1%であった．また，ベースライン抽出パターンの適合率とイベント名候補文字列抽出文字数を Table 6 に示す．ベースライン抽出パターンの網羅率は 9.4%であった．また，対象イベント名 29 件のうち，8 件の抽出パターンを用いて抽出できたイベント名とできなかったイベント名を Table 7 に示す．また，◎のついたイベントは,「ことさが」と「eventcast」の両サイトに掲載されていなかったイベントを，○はどちらか一方のサイトに掲載されていなかったイベントを表す．29 件の対象イベント名の内，抽出できたイベント件数は 13 件であった．ベースラインでは 29 件の対象イベント名の内，19 件のイベントを抽出できた．

5 考察

5.1 実験結果の考察

実験の結果，適合率は 97.1%を得た．これは 3.3 節の調査結果の,「イベント名は鉤括弧に囲まれて書かれやすい」,「“開催” が多く用いられる」を用いたことが有効であったことを示している．鉤括弧で囲まれていることで，どこまでがイベント名なのか簡単に知ることができる．しかし，ベースライン抽出パターンの結果の通り，鉤括弧のみを用いたパターンだと適合率は低い．今回は ” 開催 ” を用いたが，パターン中にイベント名に係り易い表現を含むと適合率が高くなることがわかった．まだ他にも有効な単語がある可能性があり，さらなる調査が必要である．

また，網羅率は 3.2%と低かった．これは，今回作成した抽出パターンの記事に対する一般性が低かったことを示している．しかし，イベント全体に対する網羅率は 29 件中 13 件取得出来たので 44.8%であった．また，記事数を各イベント 1,000 件に拡大して抽出を行った所，29 件中 24 件 (82.7%) のイベントが抽出できた．Table 7 中の網掛けカラムが，1,000 件に拡大して抽出できなかったイベントである．よって，今回選定した抽出パターンがイベントに対して一般性を持つことがわかる．しかし，今回選定したパターンでは，鉤括弧に囲われていないイベント名や，“開催” を含んで書かれないイベント名はすべて抽出することができない．さらなる抽出パターンの追加や鉤括弧を用いない抽出

Table 5 8 件の抽出パターン一覧とその抽出数と適合率

前部	後部	抽出数	適合率
「	」が開催	19	100%
開催中の「	」で	1	100%
で開催中の「	」	11	100%
で開催されている「	」に	0	-
開催される「	」	6	83.3%
「	」の開催	0	-
で開催された「	」	8	100%
開催されている「	」	2	100%
合計		47	97.1%

Table 6 ベースライン抽出パターンの抽出数と適合率

前部	後部	抽出数	適合率
「	」	14,972	0.9%

Table 7　抽出されたイベント名と抽出されなかったイベント名

抽出されたイベント名	抽出されなかったイベント名
◎ B-1 グランプリ	◎ PUNKSPRING
○ どきどきフリーマーケット	LIVE STAND
◎ ふくろ祭り	東京ゲームショウ
○ アフリカン・フェア	○ WBC
アンドリュー・ワイエス展	○ さっぽろ雪祭り
○ グッドデザインエキスポ	デザイン・フェスタ
スタジオジブリ・レイアウト展	○ GEISAI
◎ ベトナム・フェスティバル	METAMORPHOSE
◎ 伊賀上野ＮＩＮＪＡフェスタ	OSAKA 光のルネサンス
巨匠ピカソ展	神戸ルミナリエ
国際福祉機器展	土浦全国花火競技大会
東京おもちゃショー	東京国際映画祭
○ 北海道神宮例祭	横浜トリエンナーレ
-	◎ 湘南祭
-	WIRE
-	バリューコマース EXPO

パターンの作成を行い，記事に対する網羅率をさらに高める必要がある．

5.2　今後の展開

ここでは，今回提案した手法の将来展開について述べる．

(1) 時間表現パターンによるイベント開催判定

パターンによっては，イベントが開催中なのか，開催される予定なのか推定出来る．例えば“開催中の「　／　」”で抽出されたイベントがあれば，そのイベントはその記事が書かれた日には開催中である可能性が高い．ブログには投稿日情報がある．投稿日に対して，まだ終了していないイベントを知ることが出来る．

(2) キーワードによるイベント検索

パターンと一緒にキーワードを指定して得られた記事に対してイベント名抽出することで，キーワードに関連するイベントを取得できる．あるキーワードと特定のイベントが複数のブログ記事に共起して出現する場合は，キーワードに関連するイベントである可能性がさらに高くなる．

(3) 関連イベント検索

パターンと一緒にイベント名を指定して得られた記事に対してイベント名抽出することで，あるイベントに関連するイベントを知ることが出来る．また，あるイベントを記事に含んでいるブログサイトの全記事を対象にすることで，あるイベントに興味を持った人が他の興味を持ったイベントを知ることが出来る．

6　まとめ

本研究では，イベント推薦のためのイベント名抽出方法を提案した，今後は，提案手法のさらなる研究と，イベント情報の推薦に向けたイベント情報抽出・推薦手法の研究を行う．

付録

(A) 文字列カーネル

文字列カーネルについて説明する．

ある対象文字列に対する候補文字列の類似性を部分マッチングで評価するだけでは，表記ゆれなどの微妙な違いを考慮できない．Lodhi らが提案した文字列カーネル[2]は，文字間の距離に重みを付けた（$\lambda^{\text{距離}}(0 < \lambda \leq 1)$）一致度で類似度を計る．対象文字列 x と候補文字列 y の類似度を求める式は以下のようになっている．

$$\text{類似度 (x,y)} = \frac{k(x,y)}{\sqrt{k(x,x) \cdot k(y,y)}} \tag{3}$$

例として，“cat(x)”と“car(y)”の文字列を計る場合を説明する．まず，x に出現する文字の文字間距離，y に出現する文字の文字間距離を計算する．

“cat”であれば，“c-a”，“c-t”，“a-t”の距離を計算する．“cat”における“c-a”の距離は 2，“c-t”の距離は 3，“a-t”の距離は 2 となる．すると，式中の $k(x,x)$ は，

$$k(x,x) = (\lambda^2)^2 + (\lambda^3)^2 + (\lambda^2)^2 = 2\lambda^4 + \lambda^6$$

となる．同様に“car”における各文字間の距離を計算すると，$k(y,y)$ は，

$$k(y,y) = 2\lambda^4 + \lambda^6$$

となる．

また，$k(x,y)$ は，文字列 x，y に共通する文字

の文字間距離となる．文字列 x，y に共通して出現する文字列は “c” と “a” である．“cat” における “c-a” の距離は 2，“car” における “c-a” の距離は 2 である．よって $k(x,y)$ は，

$$k(x,y) = (\lambda^2)^2 = \lambda^4$$

となる．これら $k(x,x)$，$k(y,y)$，$k(x,y)$ を (3) の式に代入する．重み付け関数 λ に 0.9 を指定すると，“cat” と “car” の類似度は 0.35 となる．調査では，λ に 0.9 を指定し計算した．

(B)Kiwi アルゴリズム

Kiwi アルゴリズムについて，Fig.2 の Kiwi アルゴリズム概要図を用いて説明する．

Kiwi アルゴリズムとは，ある文字列に後続する文字種類数は定型的表現中には減少するという法則を用いた定型的表現抽出アルゴリズムである．例えば，Fig.2 では，「犬も」の後続文字として，「猫」「歩」「好」の 3 種類があり，文字種類数は 3 となる．後続する文字種類数が減少し続ける間は意味的切れ目でないと判断し，増加した所を意味的切れ目として文字列を切り出す．例では，「犬も歩けば棒に当たる」が定型的表現として抽出される．

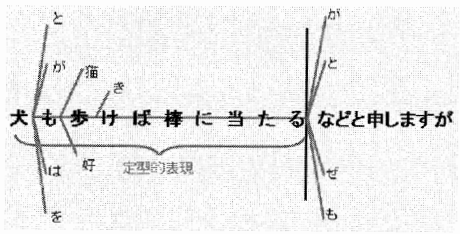

Fig. 2　Kiwi アルゴリズムの概要図

(C) 追加実験:Web 上からのイベント名抽出実験

選定した 8 件のイベント名抽出パターンを用いて Web 上よりイベント名を抽出した．抽出結果例を示す．

1. 実験手順

以下の手順で実験を行った．

(1) Yahoo!API を用いた Web 検索結果の取得

Table 8　提案手法で抽出できたイベント名の例とその開催場所

イベント名	イベントの会場
加山又三展	国立新美術館
板橋区民まつり	板橋区グリーンホール周辺
FC EXPO 2008	東京ビッグサイト
渋谷 PEACE 祭 2008	代々木公園他
ポケモンフェスタ 2008	パシフィコ横浜他

(2) イベント名抽出パターンを用いたイベント名候補文字列抽出

以下，各手順について述べる．

(1)Yahoo!API を用いた Web 検索結果の取得

今回は，Web 検索結果の取得に Yahoo!API を用いた．検索対象はブログを含んだ Web 全体とした．

(2) イベント名抽出パターンを用いたイベント名候補文字列抽出

選定した各 8 件のイベント名抽出パターンの前部と後部をフレーズとしたクエリを AND 検索で Yahoo!API に送信する．得られた検索結果のページ 1,000 件よりイベント名抽出パターンを用いてイベント名候補文字列を取得する．

2. 実験結果

Table 8 に，8 件の抽出パターンで抽出したイベント名の内，第 2 節で触れた「eventcast」と「ことさが」に掲載されていなかったイベントを開催場所と共に 5 件紹介する．また，これらはすべて去年行われたイベントである．

参考文献

1) Kumiko Tanaka-Ishii, Hiroshi Nakagawa : A Multilingual Usage Consultation Tool based on Internet Searching -More than search engine, Less than QA-, The 14th International World Wide Web Conference (WWW2005) pp.363-371. 2005.

2) H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini and C. Watkins: Text classification using string kernels, Journal of Machine Learning Research 2 (2002), pp.419-444.