blogマッピングを用いたイベント情報抽出

安村　祥子† 　池崎　正和† 　渡邉　豊英† 　牛尼　剛聡††

† 名古屋大学大学院情報科学研究科社会システム情報学専攻　〒 464–8603 名古屋市千種区不老町
†† 九州大学大学院芸術工学研究科　〒 815–8540 福岡市南区塩原 4 丁目 9–1
E-mail: †{yasumura,mikezaki,watanabe}@watanabe.ss.is.nagoya-u.ac.jp, ††ushiama@design.kyusyu-u.ac.jp

あらまし　個人の体験を時空間に関連付けて処理することは, 個人情報管理の視点から有用である. そこで, 地理情報システムでイベントを扱うことを目的に,「個人が体験し, 発生期間が限定された出来事」とイベントを定義し, イベントを処理できる地理情報システムの構築を目指す. 本稿ではそのために Web からイベント情報を自動的に抽出する手法を提案する. 本研究では, 実世界でイベントが発生したとき, Web にイベントに関する blog エントリが作成されるとし, blog エントリを収集する. そして, blog エントリからイベント情報を抽出する. イベント情報は, イベントの発生時間, および発生場所とする. しかし, イベントの発生場所を抽出する際には, 異なる領域に対応する同じ地名の存在による誤抽出, およびイベントと関係のない地名の誤抽出の問題がある. 前者の問題には地名の地理的包含関係を考慮した地名の登録により対処する. 後者の問題にはイベントを同定し, イベントの発生場所を絞り込むことで対処する. 本研究では blog エントリからイベントの発生時間, および発生場所を抽出することを目指す.

キーワード　イベント, blog, マッピング

Extraction of Information of Events by Using Blog Mapping

Shoko YASUMURA†, Masakazu IKEZAKI†, Toyohide WATANABE†, and Taketoshi USHIAMA††

† Department of Systems and Social Informatics, Graduate School of Information Science, Nagoya University
†† Faculty of Design, Kyushu University
E-mail: †{yasumura,mikezaki,watanabe}@watanabe.ss.is.nagoya-u.ac.jp, ††ushiama@design.kyusyu-u.ac.jp

Abstract　It is useful to handle the personal experiences by associating with the space and time, from a view point of managing personal information. We define an event as the occurrence that someone experiences in the real world and design a geographic information system which can handle the event. In this paper, we propose a method to extract information of events automatically from Web. We assume that when an event is occurred in the real world, blog entries about the event are registered in Web. We collect these blog entries from Web. Then, we extract information about events from blog entries. Information about the event is a time and place where the event occured. However, there are problems due to two factors. One is same place names associated with different regions. The other is place names which are not related to the event. We handle the former problem by enrollment of plece names based on geographic inclusive relation. We handle the latter problem by narrowing down place names to a place name of the event by identification of the event. In this research, we aim at extracting a time and place of the event from blog entries.

Key words　event, blog, mapping

1. はじめに

近年，写真や動画，文章など，様々なメディアデータを地図上にマッピングし, 空間上で管理するサービスが登場している. しかし，個人の生成したメディアデータを管理するならば, キーワードや空間参照のみでは不足である. 個人が体験した, 複数のメディアコンテンツに共通する事象をメディアデータと関連付け, そのコンテキストとともに管理・処理することが必要不可欠である [1]. 我々は，コンサートや展覧会のような, 個人が体験する期間が限定される出来事をイベントと定義し, イベントを処理できる地理情報システムの構築を目指す.

地理情報システムでメディアデータを管理する試みはいくつか提案されている. Sara らは，地理情報システムを用いたニュース記事の視覚的な探索を提案している [2]. しかし, このモデルでは，個人メディアデータ管理や，メディアデータ間の関連による情報探索は実現できない. 地理情報システムで個人が体験

したイベントを扱うことで, 個人の視点を反映した個人メディアデータを地図上で管理・処理できる. 地図上で処理することで, 何処で, どのようなイベントが発生したのかが視覚的に分かる. 知らなかったイベントを知り, 興味を抱いて調べれば自身の知識も広がる. さらに, イベント間の関係を表現することで [3], イベントを中心としたメディアデータの処理が実現できる.

本稿では，個人利用のためのイベント指向地理情報管理に向け，イベント情報の WWW からの自動取得を目指す. 近年, WWW の発展と大衆化により, 大量の情報ソースが WWW 上に存在する. WWW と地理情報システムに関連する研究として, WWW を地理情報システムにより拡張した, 拡張 Web 空間とその検索言語が研究されている [4]. この研究では, 地域情報サービスに向けて, ホームページと地理オブジェクトを対応付けることを目的としている. 本研究では, イベント指向地理情報管理に向けたイベント情報を Web ページから抽出する.

イベント情報の抽出対象として, 個人の情報発信手段としての blog に注目する. blog とは,「ウェブログ（weblog)」を略した言葉であり,「Web 上に残される記録」という意味をもつ [5]. blog に関する研究は数多くなされてきている. 郡らは blog から作成者の行動時の経路とその文脈を抽出し, 地図上にマッピングすることで集約して提示するシステムを提案した [6]. 倉島らは場所に関する blog から人々が旅の目的としている「対象」と,「体験」を抽出し, 地図上で提示するインタフェースを提案した [7]. 地図から探す blog 検索エンジンも提案されている [8]. これらの研究では, イベントの体験に注目しておらず, イベントと関係のない体験が対象に含まれている.

実世界でイベントが発生したとき, イベントを体験した個人により, Web 上にイベントに関する一日分の記事（以下, blog エントリ）が作成される. 本研究では, イベントに関する blog エントリを収集し, イベント情報を抽出する. イベント情報の属性として, イベントの種別, 発生時間, 発生場所などを考える. イベントの種別は抽出が困難であるため, 与えられるものとする. Web ページからの情報抽出については, これまで多くの研究がなされてきた [9]. しかし、Web ページからの情報抽出は, 属性の多様性により困難化する. そこで, イベント情報をイベントの発生時間と発生場所の情報とする. 本研究の目的は, イベント指向の地理情報システムに向けて，blog エントリからイベント情報を抽出することである.

以下, 2. ではシステムの処理全体の概要を述べる. 次に 3. では blog の収集とその内容の識別, およびイベント情報抽出処理について述べる. 4. ではイベント情報の抽出実験結果について考察する. 5. では本研究についてまとめ, 今後の課題について述べる.

2. 概　　要

2.1 イベント

イベントを，個人が体験し, 発生期間が限定される実世界での出来事と定義する. 発生期間は，多くの個人が時間を共有できる程度の期間とする. 本研究では，まず，扱うイベントを決定する．そしてイベントの種別を決定し，入力として与えるとする. 表 1 はイベントの種別の例である.

表 1　イベントの例
Table 1　Example of events

コンサート	展覧会
花火大会	お祭り
フリーマーケット	バーゲン
試合	試験
ディナーショー	試写会

2.2 blog からのイベント情報抽出

地理情報システムでイベントを扱うには, イベント情報を取得することが必要である. そこで, 実世界でイベントが発生したとき, Web にはイベントを体験した個人の, イベントの体験日記である blog エントリが作成されると仮定する. 一般に, blog エントリの HTML ソースには, メタデータが記述されている. 本研究では, イベントに関する blog エントリを収集し, 収集した blog エントリからメタデータを利用してイベント情報を抽出する.

一般に, blog エントリの内容は個人の日記から世界へ向けたメッセージなど様々である. そのような blog エントリのなかからイベントの体験日記である blog エントリを選び, 収集しなくてはならない. そのため, blog エントリを収集する際には blog エントリの内容識別が必要である. 識別後, blog エントリからイベント情報を抽出する.

抽出するイベント情報は, イベントの発生時間, および発生場所である. イベントの発生時間は, blog エントリが作成された時間とほぼ等しいとする. イベントの発生場所は, 場所を表す地名文字列とする. 例えば, 県名, 市町村名, 建築物名などである. しかし, blog エントリからイベントの発生場所を抽出する際には問題がある. まず，blog エントリに記載されている地名が複数の場所に対応する場合がある. 例えば, 緑区は名古屋市，横浜市，さいたま市，千葉市にある.「緑区」を blog エントリから抽出するのみではどの市にある緑区かが分からない. そこで本稿では, 地名の地理的包含関係を考慮した地名の登録により対処する. また，blog エントリにイベントと関係のない地名が記載されている場合がある. 例えば, 作成者がイベントの発生場所に到達するまでの通過点の地名が記載されている場合がある. その場合, 通過点の地名も抽出されてしまい, イベントの発生場所が定まらない.

同じ内容のイベントに関する blog エントリには, イベントの発生場所が通過点の地名よりも記載されやすい. そこで, 同じ内容のイベントに関する blog エントリには, イベントの発生場所が共通して記載されると考える. イベントの内容は, イベントの種別，および発生時間によりある程度同定される. 同じ種別, 同じ発生時間の, 内容が異なるイベントに関する blog エントリがある場合でも, 共通して抽出される地名が異なると考える. 例えば, 保育園での園児たちによる合奏と多目的ホールでのプロによる演奏が同じ時間にあったとする. この 2 つのイベントの種別は,「コンサート」である. しかし, これらのイベントに関する blog エントリから抽出される地名は, 前者は保育園の名称で

図 1　処理の流れ
Fig. 1　Flowchart

あり, 後者は多目的ホールの名称であり，異なる. ゆえに，同じ内容のイベントに関する blog エントリが多く収集されるほど, イベントの発生場所を正確に絞り込むことができる. 本研究では, この考えに基づきイベントの発生場所を絞り込む.

2.3　処理の概略

与えられたイベントの種別を検索語として, 個人のイベントの体験日記である blog エントリを収集する. その際, blog エントリの内容は様々であり, イベントの体験日記ではない blog エントリも収集されるため, イベント情報を抽出する前に blog エントリの内容を識別する. その後, blog エントリからイベント情報としてイベントの発生時間, および発生場所を抽出する. 日付抽出では, イベントの発生時間と blog エントリが作成された時間がほぼ等しいとして, blog エントリの最終更新日から日付を抽出する. 地名抽出では, 地名を blog エントリから検索して抽出する. 地名は, あらかじめ地名の地理的包含関係を考慮して登録されている. 地名の地理的包含関係により地名の誤抽出の問題に対処する. 抽出された地名には, イベントと関係のない地名も含まれる. そのため，同じ日付の blog エントリはイベントの発生場所に対応付けられやすいとして, イベントの発生場所を絞り込む. 図 1 は, 一連の処理の流れである.

3. 手　　法

3.1　blog 収 集

blog エントリの収集には, Google[10] を利用する. 与えられたイベントの種別，および「blog」を検索語として検索し, 検出される blog エントリを収集する. イベントの種別はイベントの定義に基づき決定されるとする. イベントの種別は blog エントリを収集する検索語であり, イベントの種別を表す複数の単語の組み合わせによる検索ができる. 例えば, 野球の試合に関する blog エントリを収集したい場合は, イベントの種別を「野球+試合」とする.

次に, 収集した blog エントリの本文箇所を抽出する. 一般的に，blog エントリの本文以外の箇所には，広告や作成者のプロフィールなど, イベントと関係のない情報が記載されているためである. blog エントリの HTML ソースにメタデータがあり, 本文の先頭箇所が記載されている場合は, それにより blog エントリの本文箇所を抽出する. メタデータがない, またはメタデータに本文の先頭箇所が記載されていない場合は, blog エントリの HTML ソースから HTML タグを除去する. さらに, 全角・半角のスペース, およびタブを除去する. それに加えて, blog の記述に一般的に使用されやすい 56 個の単語を除去する. 表 2 は, 除去対象となる単語である. これらの単語は研究の過程で収集した. blog の記述に一般的に使用されやすい単語は使用回数が大きくなる. さらに, 全ての blog エントリの記述に必ず使用されるとは限らないため, 後述する blog エントリの内容の識別に影響を与えてしまう. ゆえに, これらの単語を除去する.

表 2　除去対象となる単語
Table 2　Target words

ブログ	更新	プロフィール	ランキング
ブックマーク	参加	マイリスト	読む
追加	テーマ	タイトル	一覧
ゲスト	ユーザ	ログイン	登録
ヘルプ	ホーム	最近	最新
バックナンバー	アーカイブ	全て	表示
カテゴリ	エントリ	名前	確認
送信	投稿者	投稿時間	返事
トラックバック	コメント	メッセージ	メール
アドレス	ココログ	固定	リンク
リング	アクセス	サイト	ダイアリー
日記	記事	情報	お知らせ
こちら	コチラ	キーワード	検索
ツールバー	カレンダー	トップ	ホームページ

3.2　blog 内容の識別

blog エントリの内容を識別する前に, 識別に用いるイベントプロパティを作成する. イベントプロパティは, イベントの種別を検索語として収集された blog エントリの本文箇所の記述に使用されやすい単語と, 単語の品詞, および単語の重みの組集合である. イベントプロパティを作成するために, まず, イベントの種別を検索語として blog エントリを収集し, blog エントリの本文箇所を抽出する. 次に, 形態素解析ツール「茶筅」[11] により, 収集された全ての blog エントリの本文箇所の形態素を解析する. 形態素解析により, 名詞と, 形容詞, および自立動詞を収集する. 研究の過程で, イベントの体験日記である blog エントリの本文箇所を形態素解析し, 単語と品詞を調べた結果, この 3 種の品詞に分類される単語にイベントの特徴が表れると考えたためである. 形態素解析により単語を収集し, 単語の使用回数, および単語が抽出された blog エントリ数を算出する. 単語の使用回数は, blog エントリの本文箇所から単語が抽出された回数である. 例えば, 単語 x_i が blog エントリ A, B, C の本文箇所からそれぞれ 1, 2, 3 回抽出された場合, 単語 i の使用回数は $1+2+3$ より 6 回である. 単語が抽出された blog エントリ数は, 1 つの blog エントリの本文箇所の記述に何度使用されたとしても, 1 つとして数える. 収集された単語のうち, 使用回数が多い上位 1000 語を求める. およそ上位 1000 語以下の単語の使用回数は 1 桁と少ない. 単語が重み付けられたとしても, 重みの大きさは無視できる程度である. 1000 語それぞれに対し, 式 (1) で表される tfidf[12] により重み付ける. 式 (1) 内の $freq_i$ は単語 x_i の使用回数, $\sum_{k=1}^{1000} freq_k$ は単語の使用回数の総和，$blognum_i$ は単語

表 3　コンサートのイベントプロパティ
Table 3　Event property of coucert

し	動詞-自立	0.00265339748859604
音楽	名詞-一般	0.00593472968681857
曲	名詞-一般	0.006356380797590985
ある	動詞-自立	0.004746202189676982
いい	形容詞-自立	0.0043857912219615915
演奏	名詞-サ変接続	0.005082159724215404
思っ	動詞-自立	0.0044445263289803584
行っ	動詞-自立	0.004107331500413937
公演	名詞-サ変接続	0.006611147298315277
歌	名詞-一般	0.0048437566119204165

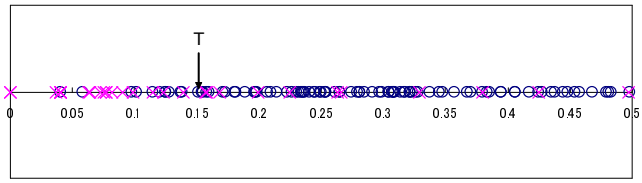

図 2　イベントプロパティによる blog の内容の識別結果
Fig. 2　Classification result of contents of blogs by an event property

x_i が使用された blog エントリ数，$blognum_{\max}$ は blog エントリ数の最大値である．

$$tfidf_i = \frac{freq_i}{\sum_{k=1}^{1000} freq_k} \times \ln \frac{blognum_{\max}}{blognum_i} \qquad (1)$$

こうして求められた, 1000 個の単語と, 単語の品詞, および重みの組集合がイベントプロパティである. 表 3 は, コンサートのイベントプロパティの一部である.

blog エントリの内容を, イベントプロパティを用いて識別する. Web からイベントの種別を検索語として収集された blog エントリの大半は, イベントの体験日記である. イベントの体験日記である blog エントリの本文箇所から抽出できる単語の使用回数は, イベントの体験日記ではない blog エントリの本文箇所から抽出できる単語の使用回数より多くなる. ゆえに, 単語の重みも大きくなる. したがって, blog エントリの本文箇所から抽出できる単語の重みを足し合わせた値が閾値を超えた場合, その blog エントリはイベントの体験日記であるとして以降処理する.

閾値は予備実験で求めた. まず，イベントの種別を「コンサート」として blog エントリを収集し，コンサートのイベントプロパティを作成した. 次に，同じイベントの種別で blog エントリを収集した. そして，blog エントリの本文箇所の記述に使用されていた単語の重みを足し合わせた値を，blog エントリの URL とともに記録した. 記録した URL により，目視で 200 件の blog エントリの内容を確認した. 図 2 は, 予備実験の結果である. T は閾値, Score は blog エントリの本文箇所の記述に使用されていた単語の重みを足し合わせた値である. ○はコンサートの体験日記である blog エントリ, ×はコンサートの体験日記ではない blog エントリを表す. 図 2 より, blog エントリ内容の識別のための閾値を, イベントの体験日記ではない blog エントリをおおむね除去できる 0.15 とした.

表 4　抽出対象となる日付の表記例
Table 4　Examples of target date

2000 年 1 月 1 日	2007.12.30	2007/01/01
2009_1_1	20070101	20071230

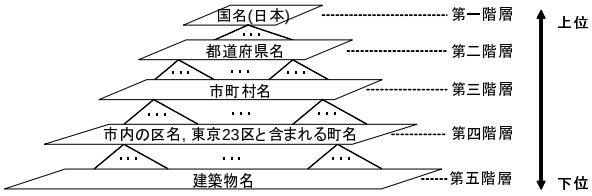

図 3　地名の階層構造
Fig. 3　Hierarchy structure of place names

3.3　日付抽出

blog エントリのメタデータ, または本文箇所から, blog エントリの最終更新日を抽出する. メタデータがない場合は, 2000 年から 2009 年の西暦の一部“ 200 ”を blog エントリの本文箇所から検索し, 日付が記載されている場合は日付を抽出する. 1900 年代は blog はあまり普及していないため抽出しない.

また, 年月は抽出するが, 日は抽出しない. イベントの発生後, イベントを体験した個人がイベントに関する blog エントリを作成するまでには, 何日かの開きがある場合がある. そこで, 日を抽出しないことにより開きをある程度無視する.

日付抽出の方法について, メタデータがある場合, ない場合に分けて述べる. メタデータがある場合は, blog の最終更新日から年月を抽出する. メタデータがない場合は, まず, blog エントリの本文箇所に“ 200 ”があるか否かを調べる. ある場合は, その次に数字があるか否かを調べる. 0 から 9 までの整数 X がある場合は, 200X を年とする. 次に, 年の後に, 7 種類の区切り文字（’_’, ’ __’, ’ 年’, ’:’, ’ : ’, ’.’）のいずれかがあるか否かを調べる. ある場合はその次に続く数字を月として抽出する. 区切り文字がない場合は, 年の後に 0 または 1 があるか否かを調べる. 年の後に 01, 02, ..., 12 のいずれかが続き, その後に日らしい数字が続く場合は日付が記述されているとし, 年月を抽出する. 表 4 は, 抽出対象となる日付の表記例である. 抽出に失敗した場合, 後述する地名の絞込みの際には 0000 年 00 月を日付として用いる.

3.4　地名抽出

blog エントリの本文箇所から地名を抽出し, 抽出された地名に blog エントリを対応付ける. 抽出する地名は, あらかじめ地名の地理的包含関係を考慮した階層構造の形式で手動で登録する. 図 3 は, 地名の階層構造の概念図である.

地理的包含関係とは, 「下位階層の地名に対応する領域が, 上位階層の地名に対応する領域に含まれる」という関係である. 例えば, 名古屋市国際展示場は港区に, 港区は名古屋市に, 名古屋市は愛知県に, 愛知県は日本に包含される. この場合, 名古屋市国際展示場は港区, 名古屋市, 愛知県, 日本に包含される. 図 4 は, 名古屋市国際展示場の場合の, 地理的包含関係の概念図である.

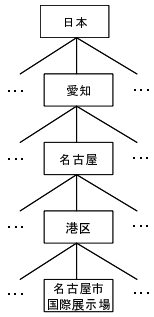

図 4　名古屋市国際展示場の地理的包含関係

Fig. 4　Geographic inclusive relation in Nagoya International Exhibition Hall

表 5　各階層における地名登録例

Table 5　Examples of enrollment of place names

第一階層	日本
第二階層	日本:愛知
第三階層	日本:愛知:名古屋
第四階層	日本:愛知:名古屋:港区
第五階層	日本:愛知:名古屋:港区:名古屋市国際展示場

階層構造は, 最上位の第一階層が日本, 第二階層が都道府県名, 第三階層が市町村名, 第四階層が市内の区の名称, および東京 23 区とそこに含まれる町名, 最下層の第五階層が建築物名で構成される. 例えば, 名古屋市国際展示場, および包含する場所の地名は, 各階層において表 5 のように登録される. そして, 最下層の地名から最上層の地名まで順に blog エントリの本文箇所から抽出される. これにより, 地名の未登録による地名の未抽出にある程度対処できる. 特に, 第五階層の建築物名には, 上位階層の地名を含むものが多い. 例えば,「名古屋市国際展示場」は「名古屋」を含む. blog エントリの本文箇所に,「名古屋市国際展示場」が記載されていた場合,「名古屋市国際展示場」が未登録であったとしても,「名古屋」を抽出できる. この場合, blog エントリを「名古屋市国際展示場」に対応付けできないが, 包含する領域の地名「名古屋」に対応付けできる.

また, 同じ場所を示す異なる地名の存在にも対処するため, 正式名称とは異なる地名も登録する. そして, 正式名称とは異なる地名が抽出された場合, 正式名称の地名に変換する. 後述する地名の絞込みにおいて, 地名に対応付けられた blog エントリ数を用いるためである. 例えば,「名古屋市国際展示場」には,「ポートメッセなごや」という愛称がある. この場合, 正式名称の「名古屋市国際展示場」, および愛称の「ポートメッセなごや」を両方とも登録する. そして,「ポートメッセなごや」が blog エントリの本文箇所から抽出された場合,「名古屋市国際展示場」に変換する.

地理的包含関係を考慮した地名登録により, 抽出された地名に対応する領域を包含する領域を示す上位階層の地名にも blog エントリを対応付けることができる. これにより, 対応する場所が異なる同じ地名の存在による, 地名の誤抽出にも対処できる. blog エントリの本文箇所から抽出された地名が複数の場所に対応する場合がある. 例えば, 緑区は, 名古屋市, 横浜市, さいたま市, 千葉市にある. ゆえに,「緑区」は 4 箇所の緑区の場所に対応する. この場合, 1 つ上の階層の地名全てについて, blog エントリの本文箇所から抽出できるかを調べる. 1 つだけ抽出された場合, その下位階層の地名として抽出する. 複数抽出された場合, または 1 つも抽出されなかった場合, どの下位階層の地名も抽出しない. 例えば,「緑区」が blog エントリの本文箇所から抽出されたとする. 緑区は, 名古屋市, 横浜市, さいたま市, 千葉市にあるため, これらの市名が blog エントリの本文箇所から抽出できるか否かを調べる.「横浜市」のみ抽出された場合, 横浜市の「緑区」として抽出する.「名古屋市」と「横浜市」が抽出された場合, どの市の「緑区」も抽出しない. どの市にある緑区かが特定できないためである. どの市名も抽出されなかった場合も同様である.

さらに, 長い地名文字列に含まれる, 短い地名文字列の地名の誤抽出にも対処する. 例えば,「東京都」には「京都」が含まれる. blog エントリの本文箇所に「東京都」が記載されていた場合,「東京都」と「京都」が抽出されてしまう. 抽出された複数の地名のなかに, 他の抽出された地名に含まれる地名がある場合, 含む地名を blog エントリの本文箇所から除去する. そして, 含まれる地名が blog エントリの本文箇所から抽出できるかを調べる. 抽出されなかった場合, その地名は blog エントリの本文箇所に記載されていなかったとみなす. 例えば,「東京都」を blog エントリの本文箇所から除去する. その後,「京都」が blog エントリの本文箇所から抽出できるかを調べる.「京都」が抽出された場合,「東京都」, および「京都」を抽出する.「京都」が抽出されなかった場合,「東京都」のみ抽出する.

これらに加えて, 人名, 特に姓からの地名の誤抽出にも対処する. 例えば,「石川某」という人名が blog エントリの本文箇所に記載されていた場合, 姓の「石川」が石川県の県名として抽出されてしまう. この場合, 形態素解析により, 人名と判定された場合は地名として抽出しない.

3.5　地名の絞込み

blog エントリの本文箇所から抽出された年月, および地名を用いて正確なイベントの発生場所を絞り込む. 地名抽出では, 地名の地理的包含関係により, 抽出された地名の上位の階層の地名にも blog エントリを対応付けた. しかし, ここでは抽出された地名のみを考慮する. 例えば, blog エントリの本文箇所から「名古屋市国際展示場」,「東京」が抽出された場合を考える.「名古屋市国際展示場」の上位階層の地名は「港区」,「名古屋」,「愛知」,「日本」である.「東京」の上位階層の地名は「日本」である. 地名抽出では「名古屋市国際展示場」, および「東京」の上位階層の, これらの地名にも対応付ける. しかし, ここでは「名古屋市国際展示場」, および「東京」の 2 つから, イベントの発生場所を絞り込む.

まず, 同年月, 抽出された地名ごとに blog エントリ数を求める. このとき, 上位階層の地名に対応付けられた blog エントリ数に, その地名の下位階層の地名に対応付けられた blog エントリ数を加える. 次に, blog エントリごとに, 抽出された地名に対応付けられた blog エントリ数を比較する. blog エントリ数が式 (2) により求められる閾値 T より大きい地名を, イベントの発生

場所として抽出する. 式 (2) 内の *mappingnum* は地名に対応付けられた blog エントリ数, *n* は抽出された地名数である.

$$T_i = \frac{3}{10} \times \sum_{k=1}^{n} mappingnum_k \qquad (2)$$

研究の過程で, 対応付けられた blog エントリ数が多い上位 3 個までの地名に, イベントの発生場所が含まれると判断した. したがって, 地名に対応付けられた blog エントリ数の総和のうち, 3 割以上が対応付けられた地名を抽出する. 例えば, 2007 年 3 月に作成された blog エントリから,「東京」,「ナゴヤドーム」,「四日市」および「大阪」が抽出されたとする. また, 同じ年月に作成された blog エントリが対応付けられた数が,「東京」が 2,「ナゴヤドーム」が 3,「四日市」が 1,「大阪」が 2 であったとする. blog エントリ数の総和は 8 である. 8 の 3 割, 2.4 がこの場合の閾値となる. 閾値を超える数の blog エントリが対応付けられた地名は,「ナゴヤドーム」である. イベントの発生場所として「ナゴヤドーム」が抽出される.

また, 地名に対応付けられた blog エントリ数が全て閾値を超えなかった場合は, blog エントリが最も多く対応付けられた地名を抽出する. 例えば, 2007 年 3 月に作成された blog エントリの本文箇所から,「東京」,「ナゴヤドーム」,「四日市」および「大阪」が抽出されたとする. 同じ年月に作成された blog エントリが対応付けられた数が,「東京」が 2,「ナゴヤドーム」が 2,「四日市」が 1,「大阪」が 2 であったとする. blog エントリ数の総和は 7 であり, 7 の 3 割は 2.1 である. 閾値 2.1 を超える数の blog エントリが対応付けられた地名はない. blog エントリ数の最大値は 2 なので, イベントの発生場所として「東京」,「ナゴヤドーム」, および「大阪」が抽出される.

日付抽出に失敗した場合は, 日付を考慮せずに地名を絞り込む. 地名ごとに対応付けられた blog エントリの数を求め, 同様に処理する.

4. 実　　験

4.1 概　　要

提案手法によるイベント情報の抽出実験をするため，Java でプロトタイプシステムを実装した. システムはイベントの種別を入力として受け取り，Google で blog エントリを収集し，イベントプロパティを作成する．その後，再び Google で blog エントリを収集し，blog エントリの内容を識別，blog エントリからイベント情報を抽出する．登録した地名数は, 全階層合わせて 4395 個である. 第一階層は 1 個，第二階層は 48 個，第三階層は 1808 個，第四階層は 1749 個，第五階層は 789 個である. 実験に向けて, 建築物名は, 多目的ホールや野球場などを主に登録した.

検索語として用いたイベントの種別は「コンサート」，および「野球+試合」である. 収集され, 処理対象とされた blog エントリのうち, イベントごとに 100 件を目視で確認した. 確認した項目は, 内容がイベントの体験日記であるか, 正確に blog エントリの最終更新日の年月が抽出されているか, イベントの発生場所が抽出され, 絞り込まれているかである.

表 6　blog 内容の識別実験結果
Table 6　Experimental result of classification of contents of blogs

	True	False
コンサート	65	35
野球+試合	48	52

表 7　日付抽出の実験結果
Table 7　Experimental result of extraction of date

	True	False
コンサート	85	15
野球+試合	90	10

4.2 blog 内容の識別実験

4.2.1 実 験 概 要

提案手法によりイベントの体験日記である blog エントリとそうでない blog エントリをどのくらいの精度で識別できるかをみるため，blog エントリの内容の識別実験をした．イベントの体験日記であると識別された blog エントリの本文箇所の記述内容がイベントの体験日記であるかをみた.

4.2.2 結果と考察

表 6 は, blog エントリの内容の識別実験結果である. True はイベントの体験日記であった blog エントリ数, False はイベントの体験日記ではなかった blog エントリ数である.

blog エントリの内容の識別実験の結果から, 確認した blog エントリのうち半数以上が個人の体験日記であったことが分かる. しかし, 検索語として用いるイベントの種別の違いにより, イベントの体験日記である blog エントリ数が大きく異なることが分かる. 収集された blog エントリ全体のうち多くを占める blog エントリの内容がイベントの種別ごとに異なっていることが考えられる. つまり,「コンサート」の場合はイベントの体験日記である blog エントリが多いが,「野球+試合」の場合はドラフト会議に関する blog エントリが多く, イベントの体験日記である blog エントリが少ないということである. 識別の閾値をイベントの種別ごとに変化させる必要がある. さらに, 適切な検索語を用いる必要がある. 若木らは質問者の期待する内容に特化した検索結果を得られるような検索語を提示し, 質問にあった曖昧性を解消するための手法を提案している [13]. イベントの種別ごとに適切な検索語を求めれば, イベントの体験日記である blog エントリをより多く収集できると考えられる.

4.3 日付抽出実験

4.3.1 実 験 概 要

提案手法により blog エントリの作成された年月をどのくらいの精度で抽出できるかをみるため, 日付抽出実験をした．目視で blog エントリの最終更新日を調べ, その年月が抽出された年月と同じであるか否かをみた.

4.3.2 結果と考察

表 7 は, 日付抽出の実験結果である. True は blog エントリが作成された年月の抽出に成功した blog エントリ数, False は blog エントリが作成された年月の抽出に失敗した blog エントリ数である.

表 8　地名抽出および地名の絞込み実験結果
Table 8　Experimental result of extraction and narrowing down place names

	Before/After	True	Unknown	False
コンサート	Before	39	19	42
	After	39	5	56
野球+試合	Before	5	27	68
	After	6	6	88

日付抽出の結果から, イベントの種別の違いによらず, 高い精度で年月を抽出できたことが分かる. メタデータがある blog エントリが多いことに加え, メタデータがない blog エントリに対する本稿の日付抽出手法が適切であったといえる. 日付抽出に失敗した場合の大半は HTML タグ内に日付がある場合であった. blog エントリの本文箇所を抽出する際に HTML タグを除去するために失敗したのである. また, 日付の区切り文字としてスペースが用いられている場合もあった. スペースもまた, blog エントリの本文箇所を抽出する際に除去するために, 日付抽出の際に日付の区切り文字として使用できなかった. 例えば, 日付が “2007 1 1” と記述されていた場合, スペースを除去すると “200711” となる. 提案手法では, “20070101” と記述されていなければ抽出しないため失敗した. さらに, 最終更新日の前にそれとは異なる日付が記載されている場合もあった. 日付を 1 つ抽出すると, 他に記載されている日付は抽出しないため抽出に失敗した. blog エントリの本文箇所に記載されている全ての日付を抽出し, 最も新しい日付を求める必要がある.

4.4　地名抽出および地名の絞込み実験

4.4.1　実 験 概 要

提案手法によりイベントの発生場所をどのくらいの精度で抽出できるかをみるため, 地名抽出および地名の絞込み実験をした. 抽出された地名が blog エントリにイベントの発生場所として記載されているかをみた.

4.4.2　結果と考察

表 8 は地名抽出および地名の絞込み実験の結果である. Before/After は地名抽出後に, 地名を絞り込む前後を表し, True はイベントの発生場所のみが抽出された blog エントリ数, Unknown はイベントの発生場所と, イベントと関係のない地名が抽出された blog エントリ数, False はイベントの発生場所が抽出されなかった blog エントリ数である.

地名抽出の結果から, 検索語として用いるイベントの種別の違いにより, イベントの発生場所を抽出できた blog エントリ数が大きく異なることが分かる. 地名抽出に失敗する場合は, 次の 3 通りに分けられる. blog エントリにイベントの発生場所が記載されていない場合, 地名が記載されていても, その地名が登録されていない場合, 地名文字列を含む, 地名と関係のない記述がある場合である. 例えば, 個人的な小規模なコンサートの場合, イベントの発生場所が記載されていない, 記載されていてもその地名が登録されていないことが多い. 大規模なコンサートの場合でも, 楽団名に地名文字列が含まれているためにイベントと関係のない地名が抽出されることがある. 野球+試合の場合は, 球団名に地名文字列が含まれているためにイベントと関係のない地名が抽出されることが多い. また, 「○○選手」の姓の部分「○○」が登録されている地名に含まれている場合, 地名として抽出されることもある. 人名であった場合に対する対処が不完全であったといえる. この問題に対処するには, イベントの種別ごとに, 地名抽出をする前に blog エントリの本文箇所から除去する単語を収集する必要がある. 例えば, コンサートならば楽団名など, 野球+試合ならば球団名, 選手名などである.

地名の絞込みの結果から, イベントの発生場所が除去され, イベントと関係のない地名が抽出される場合が多いことが分かる. 原因として, 収集した blog エントリ数の不足が考えられる. 正確なイベントの発生場所へ対応付けられた blog エントリ数が不足したのである. イベントを体験した人々全員が必ずイベントに関する blog エントリを作成するとは限らない. 作成したとしても, 全ての blog エントリを収集できるとは限らない.

また, 日付抽出における問題もある. 日付抽出では, イベントの発生時間はイベントに関する blog エントリの最終更新日とほぼ等しいとして, blog エントリの最終更新日を抽出した. しかし, 例えば, 月末にイベントが発生した場合, その月とその次の月にイベントに関する blog エントリが作成される時間がまたがってしまう場合がある. イベントが発生した数年後の, イベントが発生した日と同じ日に, 数年前のイベントに関する blog エントリが作成されるかもしれない. そのような場合には対応できない. 本稿の提案手法をより適切に用いるには, イベントの発生時間をより厳密に抽出する必要がある.

さらに, イベントを体験した人々全員がイベントに関する blog エントリを作成し, イベントの発生時間が正確に抽出されるとしても問題がある. 同じ時間に複数のイベントが発生し, 抽出された地名の階層が異なる場合である. 例えば, 東京に住む人がナゴヤドームでのイベントを体験した場合を考える. その人は, 東京での生活, およびナゴヤドームで体験したイベントに関する blog エントリを作成する. 地名抽出において, blog エントリから「東京」, および「ナゴヤドーム」が抽出されたとする. ナゴヤドームでのイベントと同じ発生時間に, 東京ドームでイベントが発生したとする. 同様に, 東京ドームでイベントを体験した人々が, そのイベントに関する blog エントリを作成する. 「東京ドーム」を含む,「東京」の下位階層の地名に対応付けられた blog エントリ数は,「東京」に対応付けられた blog エントリ数に足し合わされる. 「東京」に対応付けられた blog エントリ数が「ナゴヤドーム」に対応付けられた blog エントリ数を大きく上回る場合, 地名の絞込みにおいて,「東京」がイベントの発生場所として抽出されてしまう. 図 5 に, 地名の絞込みの失敗例を示す.

この問題への対処として, 下位階層の地名が抽出された場合はそれより上位階層の地名は抽出しないこと, 同じ階層の地名が複数抽出された場合は, それぞれの地名の blog エントリ数を比較することが考えられる. この例の場合,「ナゴヤドーム」は「東京」より下位階層の地名であるため,「ナゴヤドーム」のみが抽出される. しかし, その場合, 地名の未登録による地名の未抽出の問題には対処できない. より多くの地名, 特に建築物名を登録する必要がある. さらに, より多くのイベントの体験日記で

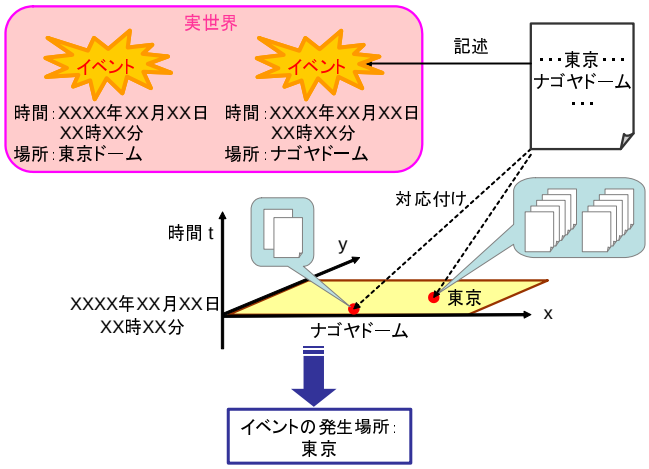

図 5　地名の絞込みの失敗例

Fig. 5　Failure example of narrowing down place names

ある blog エントリを収集する必要がある. イベントと関係のない blog エントリにはイベントと関係のない地名が記載されている. イベントと関係のない地名が建築物名であれば必ず抽出されてしまう. ゆえに, イベントの体験日記である blog エントリが多いほどイベントの発生場所を絞り込める.

5. おわりに

5.1 まとめ

我々はイベントの体験日記としての blog エントリに注目し, blog エントリからのイベント情報抽出手法を提案した. 提案手法では, 与えられたイベントの種別, および抽出したイベント情報によりイベントをある程度同定することでイベントの発生場所を絞り込んだ. 提案手法に基づきプロトタイプシステムを実装し, 実験をした. その結果, イベントの種別によるが, blog エントリ内容の識別により, イベントの体験日記である blog エントリを収集した. 日付抽出において, blog エントリの最終更新日を高い精度で抽出できた. しかし, イベントの発生場所の抽出において, 満足な結果を得られなかった. 実験により, 登録した地名数が不足しているために地名を抽出できないこと, blog エントリの本文箇所に記載されている地名には, イベントと関係のない地名が多いことが分かった. また, イベントの発生場所が blog エントリの本文箇所に記載されていない場合が予想以上に多いことも分かった. 地名を絞り込むとしても, 提案手法では, 地名抽出においてイベントの発生場所が抽出されていなければならない.

5.2 今後の課題

今後の課題は, 収集する blog エントリ数の増加と, より正確なイベント情報を抽出できるようにすることである. 同じイベントに関する blog エントリ集合を求めることができれば, イベント情報の抽出精度を高める方法もある. また, イベントの発生場所の地名を抽出するために, さらなる地名の登録が必要である. さらに, イベント情報を地理情報システムで利用するには, イベントの発生場所と地図上の座標を対応付ける必要がある. 建築物名を含む地名を入力することで, 対応する座標を得られるサービスがある [14][15]. これらのサービスは一部の建築物名の入力に対応していないため, 本研究ではこれらを利用しなかった. 地名と地図上での座標を対応付けることも重要な課題である.

謝辞

本研究の一部は大幸財団の研究助成によって実施された.

文　献

[1] 牛尼剛聡, 利用者の経験に基づいた個人コンテンツ検索・推薦のモデル, DBSJ Letters, Vol.5, No.1, pp.77-80, 2006.

[2] Sara Irina Fabrikant, Visualizing Region and Scale in Information Spaces, Proc. of The 20th International Cartographic Conference(ICC 2001), Beijing, China, Aug.6-10, 2001, pp.2522-2529, 2001.

[3] Masakazu Ikezaki, Event Handling Mechanism for Retrieving Spatio-temporal Changes at Various Detailed Level, Proc of IEA/AIE 2005, pp.353-356, 2005.

[4] 平松薫, 地域情報サービスのための拡張 Web 空間, 情報処理学会論文誌, Vol.41, No.SIG6(TOD7), pp.81-90, 2000.

[5] Blog とは?, http://blog.goo.ne.jp/info/bloginfo1.html

[6] 郡宏志, ブログからのビジターの代表的な行動経路とそのコンテキストの抽出, 信学技報, DE2006-55(2006-7), pp.29-34, 2006.

[7] 倉島健, Blog からの街の話題抽出手法の提案, 電子情報通信学会第 16 回データ工学ワークショップ (DEWS2005 2C-i10), 2005.

[8] maplog, http://maplog.jp/

[9] Chia Hui Chang, A Survey of Web Information Extraction Systems, IEEE Transaction on Knowledge and Data Engineering, Vol.18, No.10, pp.1411-1428, 2006.

[10] Google, http://www.google.co.jp/

[11] 茶筅, http://chasen.naist.jp/hiki/ChaSen/

[12] J.K. Sparck, A statistical interpretation of term specificity and its application in retrieval, Journal of Documentation, Vol.28, No.1, pp.11-21, 1972.

[13] 若木裕美, 検索語の曖昧性解消のためのトピック指向単語抽出および単語クラスタリング, 情報処理学会論文誌, Vol.47, No.SIG19(TOD32), pp.72-85, 2006.

[14] 地図閲覧サービス, http://watchizu.gsi.go.jp/

[15] ジオコーディング, http://map.fkoji.com/geo/