D-045

地理情報システムでのイベント管理に向けた Blog 処理

Blog Management for Handling Events in GIS

安村 祥子† **池崎 正和**† **渡邉 豊英**† **牛尼 剛聡**‡
Shoko Yasumura Masakazu Ikezaki Toyohide Watanabe Taketoshi Ushiama

1．はじめに

近年, 様々なメディアデータを地図上にマッピングし, 管理する Web サービスが登場している. これらのメディアデータは, 関連付けられた場所で起きた出来事・体験を表現したものと考えられる. 我々は, 個人が体験し, 期間が限定された出来事をイベントと定義し, 個人の情報発信手段である Blog に注目する. Blog とイベントを関連付けることによる Blog 読者用イベント検索システムの構築を目指す.

これまで, 我々は Blog 集合からイベントを抽出する手法を提案してきた[1]. 本稿では提案手法によるイベント抽出実験の結果を分析し, 議論する.

以下, 2 章ではシステムの処理概要を述べる. 次に 3 章では Blog の収集, 内容の識別, およびイベント抽出処理について述べる. 4 章で, イベント抽出の実験を通して提案手法を評価し, 5 章では本稿についてまとめる.

2．概要

まず, イベントの種別（例えば, コンサート, 展覧会など）を検索語として, 一日分の Blog 記事・blog entry を検索・収集する. このとき, 収集された blog entry には, イベントと関係のない内容のものが多く含まれるため, blog entry の内容を識別する必要がある. 識別後, イベントを抽出する.

イベントが一意に特定される条件を, その発生時間, および発生場所が明らかであることとし, イベントの情報としてイベントの発生時間・発生場所を抽出する. イベントの発生時間は, blog entry が作成された時間とほぼ等しいと考え, blog entry の最終更新時間から抽出する. イベントの発生場所として地名文字列を抽出する. しかし, イベントと関係のない地名が blog entry に記載されている場合もある. イベントが特定される条件に基づき, 同じ時間・場所の情報を抽出できる blog entry 数が多い場合, その時間・場所でイベントが発生したとする. 抽出したイベントの情報により blog entry に記載されているイベントの発生場所を絞り込む.

3．手法

3.1 Blog 収集

イベントの種別, および”blog”を検索語として blog entry を収集する. 次に, blog entry の本文以外の箇所にはイベントと関係のない情報が記載されているため, HTML ソース中のメタデータに記載されている本文の先頭箇所を利用して blog entry の本文を抽出する. メタデータがない場合は, HTML タグを除去した残りを本文として利用する.

3.2 Blog 内容の識別

収集された blog entry の内容を識別するため, イベントプロパティを作成する. イベントプロパティは, blog entry の本文に記載されやすい単語, 単語の品詞, および単語の重みの組集合である. まず, blog entry を収集し, 本文を形態素解析し, 名詞, 形容詞, および自動詞を収集する. 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 entry 数を算出する. 単語の使用回数が多い上位 1000 語を求め, 式(1)で表される *tfidf*[2]により重み付ける. およそ上位 1000 語以下の単語は使用回数が小さく, その重みの大きさは無視できる程度であるため利用しない. 式(1)内の $freq_i$ は単語 x_i の使用回数, num_i は x_i が使用された blog entry 数, num_{max} は blog entry 数の最大値である. この 1000 個の単語, 品詞, および重みの組集合がイベントプロパティである.

$$tfidf_i = \frac{freq_i}{\sum_{k=1}^{1000} freq_k} \times \ln \frac{num_{\max}}{num_i} \quad (1)$$

各 blog entry の内容を, イベントプロパティを用いて識別する. blog entry の本文から抽出される単語の重みの和が閾値を超えた場合, イベントと関係のある内容であるとして以降処理する. 閾値は予備実験で求めた.

3.3 日付抽出

blog entry のメタデータ, または本文から blog entry が作成された年月を抽出する. イベント発生から, イベントを体験した個人がイベントに関する blog entry を作成するまでには時間差が存在する. そこで日を抽出しないことでこれに対応する. メタデータがある場合は, blog entry の最終更新時間から年月を抽出する. メタデータがない場合は, 本文に年号の一部”200”があるかを調べる. ある場合は以降に日付が記載されていると判断し, 抽出する.

3.4 地名抽出

blog entry の本文から抽出された地名に blog entry を対応付ける. 地名抽出に利用する地名辞書は, 地理的包含関係を考慮した階層構造の形式で, 手動で登録しておく. 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. 図 1 は地名の階層構造の例である.

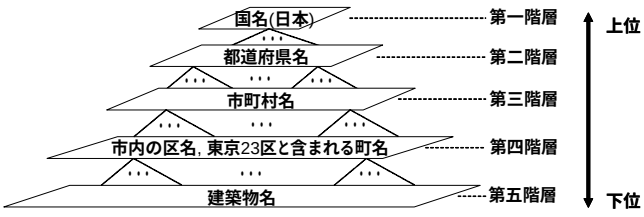

図 1 地名の階層構造

階層構造は, 最上層の第一階層が日本, 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. これにより, 抽出された地名の上位階層の地名にも blog entry を対応付けることができる.

†名古屋大学大学院 情報科学研究科
‡九州大学芸術工学研究院

3.5 地名絞込み

blog entry の本文から抽出された年月, および地名を用いてイベントの発生場所を絞り込む. まず, 同じ年月, 地名ごとに blog entry 数を求める. 次に, blog entry 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, n は抽出された地名数, i は blog entry の識別子, *mappingnum* は地名に対応付けられた blog entry 数（地名の抽出数）である.

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

対応付けられた blog entry 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考えたため, 係数を 3/10 とした. また, 地名に対応付けられた blog entry 数が全て閾値を超えなかった場合は, blog entry 数が最大の地名を抽出する. 日付抽出において年月の抽出に失敗した場合は年月を考慮せずに地名を絞り込む.

4．実験

4.1 実験概要

提案手法に基づき Java でプロトタイプシステムを実装し, イベント抽出実験をした. システムはイベントの種別を入力とし, Google で blog entry 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 entry を収集し, blog entry の内容を識別し, blog entry からイベントを抽出する. イベントの種別は「コンサート」, 「野球+試合」とした. 実験に向けて, 全階層合わせて 4395 個の地名を登録した. 建築物名は多目的ホールや野球場などである. 収集され, 処理対象とされた blog entry のうち, イベントの種別ごとに次の 2 項目について, 100 件を目視で確認した.

- 内容がイベントと関係があるか
- イベントの発生場所が抽出されているか

4.2 Blog の内容の識別実験の結果と考察

表 1 は blog entry の内容の識別実験結果である. True はイベントに関係のある内容であった blog entry 数, False はイベントに関係のない内容であった blog entry 数である.

表 1 blog 内容の識別実験結果

	True	False
コンサート	65	35
野球+ 試合	48	52

結果から, True の blog entry 数がイベントの種別により大きく異なることが分かる. 野球+試合の場合, シーズンオフのためドラフト会議に関する内容が多く, 試合に関する内容が少ないためである. このように, 結果が季節により変化するイベントがあることが分かる. また, 提案手法では, イベントの種別を検索語として収集された blog entry にはイベントの体験日記が多く含まれるとして *tfidf* によりイベントプロパティを作成した. しかし, イベントの体験日記である blog entry の収集された blog entry 全体を占める割合により識別精度が変化すると考えられる. 提案手法をより適切に用いるには, イベントの種別ごとにイベントの体験日記である blog entry を手動で収集し, それらの blog entry のみを用いてイベントプロパティを作成する. そして, イベントの種別ごとに予備実験をすることで適切な閾値を設定する必要がある. さらに, 本文が長いために単語の重みを足し合わせた値が大きくなり, イベントの体験日記であると誤識別される場合があるため, 正規化する必要がある.

4.3 イベントの発生場所抽出実験の結果と考察

表 2 は地名抽出および地名の絞込み実験の結果である. Before/After は地名を絞り込む前後, True はイベントの発生場所のみが抽出された blog entry 数, Unknown はイベントの発生場所, およびイベントに関係のない地名が抽出された blog entry 数, False はイベントの発生場所が抽出されなかった blog entry 数である.

表 2 イベントの発生場所抽出実験結果

	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

結果から, イベントの発生場所がほとんど絞り込まれていないことが分かる. 同一イベントに関して, 正確なイベントの発生場所の抽出数が少なかったためである. イベントの発生場所の抽出数を大きくするため, 同一イベントに関する blog entry をより多く収集しなければならない. ゆえに, 検索語をより具体的にする必要がある. また, blog entry の本文抽出の不正確さ, 地名文字列を含む地名と関係のない文字列の記述, イベントの発生場所が記述されていないなどの問題により, 上位階層の地名の抽出数がイベントの発生場所の抽出数より大きくなり, 正確なイベントの発生場所の同定に失敗した. イベントの発生場所の抽出精度を高めることのみを考えるならば, 対応付けられた blog entry 数が閾値以上の地名ではなく, 閾値未満の地名を抽出すべきである. しかし, それではイベントが特定される条件に従わない. 条件に基づきイベントの発生場所を抽出するには, まず, イベントと関係のない情報から地名が抽出される場合を減らす必要がある. 具体的には, イベントの種別ごとにあらかじめ地名文字列を含む地名と関係のない文字列を収集し, 地名抽出の前に本文から除去するなどが考えられる.

5．おわりに

本稿では, 提案手法によるイベント抽出実験の結果についてより具体的に議論した. それにより, 提案手法をより有効に適用する方法が明らかになった. 今後の課題は, 提案手法に改良を加え, イベントの抽出精度を高めることである. さらに, 地名と座標を対応付け, イベントを地図上に示す方法を考案することも重要である.

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.

参考文献

[1] 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, D8-3, 2007.

[2] 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.