3 ブログからの地域イベント情報抽出

岡本昌之　菊池匡晃　（(株)東芝　研究開発センター）

近年位置情報サービスが盛んになっているが，利用されているのは静的な情報である．ライブ演奏などのイベント情報や短期間の出店情報など静的でない地域イベント情報はほとんど扱われていない．これらの情報は Web 上のブログなどで短期間のうちに地名とあわせて言及される件数が増加しやすい．したがって広く世の中のホットな話題を抽出するのに利用されてきた話題抽出技術を適用することで取得できる．本稿では，地名に言及したブログエントリを入力とし，2 段階の階層クラスタリングと時事性，地域性の考慮により地域イベント情報を抽出する手法を解説する．また，東京と神奈川の 37 地域を対象とした評価実験を通じた現状の精度と課題を述べる．

ホットな話題を見つけるサービスの広がり

携帯端末の普及に伴い，ナビゲーションや情報推薦のための位置情報サービスが多数開発・公開されている．最近では，従来からのカーナビゲーションシステム向けの施設情報提供だけでなく，携帯電話やスマートフォン向けにグルメサイトやタウン情報サイトなどが提供するサービスが広く利用されている．これらのサービスで用いられるレストランや施設などの地点情報（POI：point of interest）は，登録・更新頻度の違いはあっても長期間存在することが想定された静的なものである．

世の中には，これらの登録された POI 以外のイベント会場などのイベント情報や，小規模な口コミ情報が溢れている．たとえば，Web 上のブログでは「○○で開催中のライブは最高だった」「□□駅近くのイタリア料理店☆☆のパスタはおいしかった」といった記述が多数ある．このような情報を用い，ある地域において話題となっているイベントや口コミ情報（地域イベント情報）を取り入れることで，ユーザにとってより有益な情報提供の支援につながると考えられる．

地域イベント情報の抽出には，世間でホットな話題を表すキーワードを抽出し，可視化する話題抽出技術の適用が有望と考えられる．すでにこれらの技術を用いたサービスも増えており，たとえばきざしランキング☆1，ホットワードリンク☆2 などが公開されている．話題抽出に関してはさまざまな研究が行われており，単語の重み付け，話題性の計算，文のモデリングによる手法がニュース記事やブログ記事に対して適用されている[1]．また，大規模な社会事象に対する時空間的な話題の推移を調べる手法も提案されている[5]．

しかし，特定の地域イベントについて述べた記事は世間でホットな話題と比べエントリ数が少なく，これらの方法をそのまま適用することは難しい．

我々は，1 つのアプローチとして対象とする地名等で絞り込んだ結果を情報源として用いた．たとえば，日本の都市部では，駅名をキーとしてイベント情報を抽出することができると考えられる．さらに，地域性を考慮することで出力の精度を上げることを試みた[4), 6)]．

本稿では，このような性質を利用して地名に基づく時系列テキストデータから，地域イベント情報を抽出する手法と，東京・神奈川を対象とした評価実験について述べる．

地域イベント情報の抽出

◉抽出の流れ

図 -1 に地域イベント情報抽出の処理の流れを示す．

（1）「東京」や「秋葉原」などの地名をクエリとして，一般のブログ検索エンジンを利用して地名を含むブログエントリを収集する．スパムブログ対策のため，あらかじめ指定した NG ワードを含むエントリは除去される．

（2）収集したエントリに対し形態素解析，固有表現抽出を行い，各単語を事前に生成した IDF（inversed document frequency：単語が出現する文書数の逆数）によって重み付けした文書ベクトルを生成

☆1 http://kizasi.jp/
☆2 http://tvsurf.jp/w/pc/hwl/

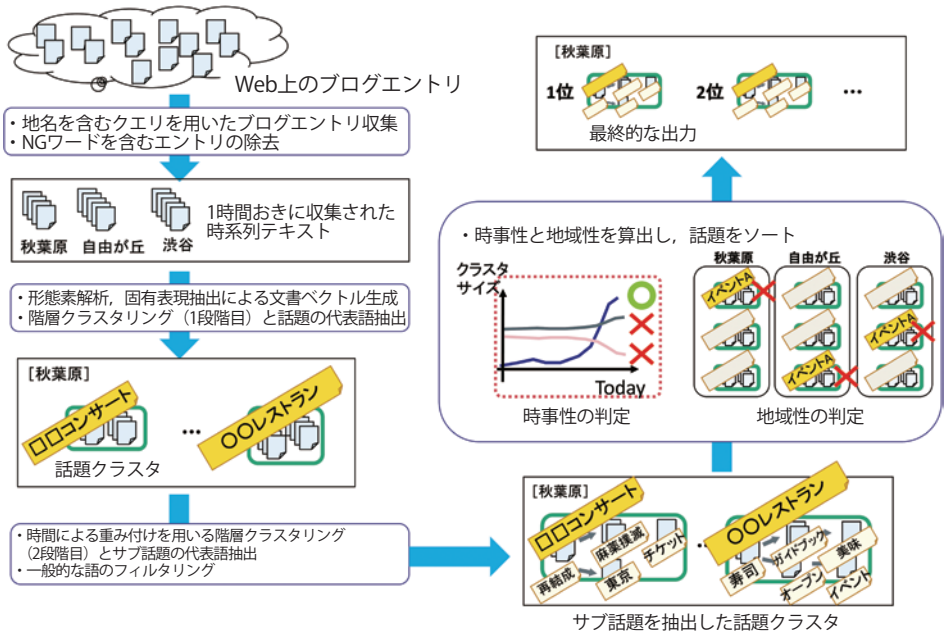

図 -1　地域イベント情報抽出の流れ

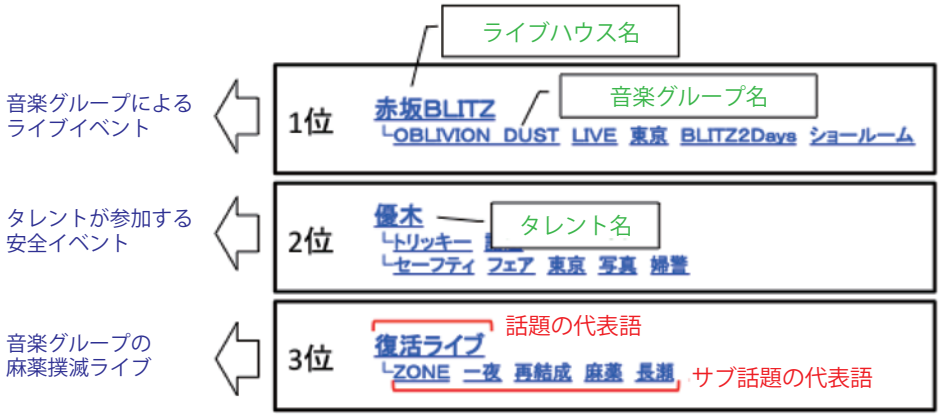

図 -2　抽出結果の例

する．そして，余弦尺度を用いた凝集型階層クラスタリング[3]を行う．分類された各クラスタからC-value 法[2]により話題の代表語が抽出される．

(3) 1 つの話題から時間の経過に伴った内容の推移を抽出するために，生起時刻の近い文書が同一のクラスタを形成しやすいよう類似度計算式を重み付けした上で，各話題クラスタを再度クラスタリングしてサブ話題と代表語が抽出される．後処理として，話題を十分に特定できない一般的な語はフィルタリングされる．

(4) 時事性と地域性の判定が行われる．時事性に関しては，直近の期間に含まれる文書数が集計期間の文書数に対してどれだけ多いかを検定し，有意に多いものをホットな話題とみなす．地域性に関しては，同じ日に別地域で，同じ話題の代表語を持つクラスタの数を算出し，その値が大きい場合は，地域性の低い話題であると判断され，除外される．

最終的に，各地域におけるホットな話題が抽出される．図 -2 に抽出された話題の例を示す．話題の代表語と他のキーワード群を用いることでモバイル端末などの小さい画面でもコンパクトに表示できる．

◉抽出されるイベントの特徴

ブログ検索結果のデータから抽出された典型的なイベントやスポットとしては以下のようなものがある．

- 音楽グループの告知されていないイベント（レミオロメン ゲリラライブ：渋谷）
- 短期間の出店情報（福島物産展：六本木）
- スポーツイベント（東京レインボーウォーク：台場）
- 展博イベント（加賀百万石名品展：池袋）
- 美術や写真の個展（原研哉デザイン展：吉祥寺）
- 有名レストランや名物メニュー（堂島ロール：銀座）

東京，横浜，渋谷，銀座，川崎，上野，鎌倉，品川，赤坂，町田，目黒，日本橋，府中，両国，新宿，池袋，六本木，秋葉原，原宿，恵比寿，台場，吉祥寺，浅草，表参道，下北沢，新橋，横須賀，代官山，有楽町，築地，自由が丘，小田原，みなとみらい，後楽園，茅ヶ崎，神楽坂，汐留

表 -1　実験対象の駅一覧

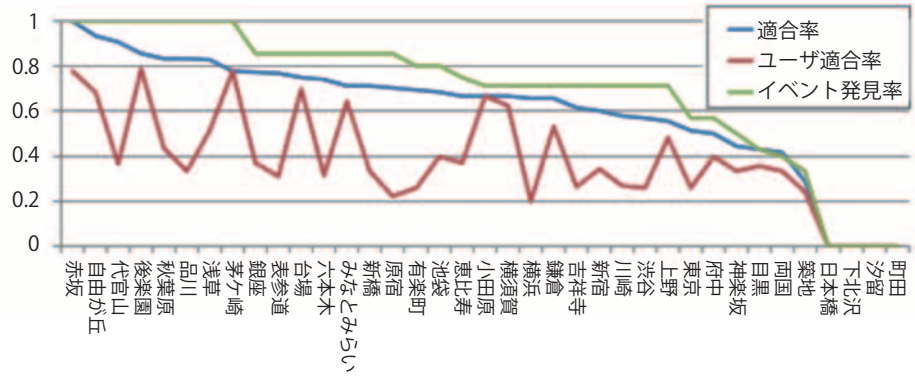

図 -3　地名ごとの適合率・ユーザ適合率・イベント発見率

上記のような例に加え，ニュースなどで報道された事件などが抽出される．

イベント抽出の評価

提案手法により地域イベントをどの程度抽出可能かを調査するため実験を行った．

◉対象データ

東京および神奈川の 822 駅を対象とした予備調査において，1 日あたりの平均エントリ数が 100 を超えていた駅のうち，表 -1 に示す 37 駅を対象として用いた．

◉抽出精度

抽出された話題のうち，どの程度が地域イベントとして認識できるかについて 1 イベントあたり 2 人の被験者が評価を行った．それぞれの日付，場所のペアにおいて上位 5 つまでの話題について以下の 3 段階で評価値を付与した．

(2)：キーワード群から地域イベントと認識できる
(1)：キーワード抽出前の元記事群から地域イベントを認識できる
(0)：地域イベントと認識できない

各被験者は，どのようなイベントの情報が含まれているか，どこでそのイベントがある，またはあったか，の 2 点に基づいてスコア付けした．

2009 年 3 月 3 日から 3 月 9 日までの 1 週間の 37 駅分のデータから抽出された 764 の話題を評価した結果の全体平均は以下の通りである．

- **イベント発見率**：77.0%
- **適合率**：66.6%
- **ユーザ適合率**：39.3%

ここで，イベント発見率は 1 つの地名について出力された最大 5 つの話題に対し，1 つ以上評価値 2 の話題が含まれる比率を表す．つまり，平均して 77% の駅に対しては何らかの地域イベントを認識できていることが分かる．

また，適合率は 1 つのイベントに対し少なくとも 1 人の被験者が 2 または 1 を付与した比率を示し，ユーザ適合率は 1 つのイベントに対し少なくとも 1 人の被験者が 2 とスコア付けした比率を表す．つまり，ユーザが元のブログの文章を読むことが許容される場合，平均して 66.6% の抽出結果は地域イベントとして認識されるが，キーワード群のみから地域イベントと認識できた割合は 40% 程度にとどまることが分かる．

また，各駅についての結果を図 -3 に示す．横軸はイベント発見率の高い順に並べた地名を表し，縦軸はイベント発見率，適合率，ユーザ適合率を表す．図 -2 より，適合率とユーザ適合率の間には相関が見られず，ユーザが地域イベント情報と認識できない場合が多いことが分

図 -4　位置情報サービスとの連携

かる．これには，大きく以下の理由が挙げられる．

- **地名の他の用例の有無**：たとえば，同じ表記でも地名「川崎」は「川崎駅」「川崎市」「川崎区」のようにさまざまな広さの地域が含まれ，また同じ表記は人名など他にもよく用いられる．したがって，地名関連以外の話題も多数抽出される．
- **地域イベントと判断するための情報不足**：会場名や建物名などの地域イベントと判断するために重要な単語がキーワード群に含まれていない場合があった．これらの名称は省略や別名など異なる表記で書かれることも多く，代表語の抽出において C-value が小さくなることが原因の 1 つである．
- **ユーザによる背景知識の違い**：たとえば，店名や施設名などは特段の規則があるわけではなく，知っている人にはすぐに分かるが，知らない人には意味が分からない場合も多い．このような個人差はユーザ適合率に大きく影響すると考えられる．

また，今回評価したイベントはあくまで被験者がイベントと認識したかどうか，という主観的なものであり，実際にイベントであるかどうか，いつ実施される／されたイベントであるかは考慮されないため，それらの評価基準の検討も重要である．

このように，現状では課題が多いものの，話題抽出技術に話題性・地域性の要素を加えることで，ある程度地域イベント情報活用の可能性を見出せたと考えられる．

今後の展望

ここまで，話題抽出技術の地域イベント情報検出への応用について述べた．従来の位置情報サービスとも組み合わせることで，たとえば図 -4 のように，近隣の店舗や施設だけでなく，より揮発的なイベントなどを知る機会を増やすことができると期待される．

本稿では，地域イベント情報の抽出の最初の手掛かりとして地名を用いたが，地理情報システムを組み合わせることで，緯度経度などを組み合わせた手法も用いることができる．

また，より即時性の高い情報を抽出する手段として，twitter[☆3] などのマイクロブログの活用は有望と考えられる．twitter は日本では活用が広がる途上であるが，筆者らの集計では，前述の評価で用いた地名の場合 7 月下旬で 1 日あたり 4000 エントリ，9 月末で 1 日あたり 7000 エントリと約 2カ月で大きく伸びている．緯度経度の活用に関しても同様に利用でき，こちらも同様にエントリ数が伸びている．

提示手段としては，AR（Augmented Reality：拡張現実感）技術も本格普及の兆しを見せており，近い将来これらを融合したサービスの展開が期待される．

☆3　http://twitter.com/

参考文献

1）Chen, K-.Y., Luesukprasert, L. and Chou, S. T.：Hot Topic Extraction Based on Timeline Analysis and Multidimensional Sentence Modeling, IEEE Trans. Knowledge and Data Engineering, No.19, Vol.8, pp.1016-1025 (2007).
2）Frantsi, K. and Ananiadou, S.：Extracting Nested Collocations, in Proc. Int. Conf. on Computational Linguistics (COLING 1996), pp.41-46 (1996).
3）Kamvar, S., Klein, D. and Manning, C.：Interpreting and Extending Classical Agglomerative Clustering Algorithms Using a Model-Based Approach, in Proc. Int. Conf. on Machine Learning (ICML2002), pp.283-290 (2002).
4）菊池匡晃，岡本昌之：ブログエントリからの地域イベント情報抽出，マルチメディア，分散，協調とモバイル (DICOMO2009) シンポジウム論文集，pp.218-225 (2009).
5）Mei, Q., Liu, C., Su, H. and Zhai, C.：A Probabilistic Approach to Spatiotemporal Theme Pattern Mining on Weblogs, in Proc. Int. World Wide Web Conference (WWW2006), pp.533-542 (2006).
6）Okamoto, M. and Kikuchi, M.：Discovering Volatile Events in Your Neighborhood：Local-Area Topic Extraction from Blog Entries, in Proc. Asia Information Retrieval Symposium (AIRS2009), LNCS 5839, pp.181-192, Springer (2009).

（平成 21 年 11 月 2 日受付）

岡本昌之（正会員）
masayuki4.okamoto@toshiba.co.jp
2003 年京都大学大学院情報学研究科社会情報学専攻博士後期課程修了．博士（情報学）．同年（株）東芝入社．現在，研究開発センター知識メディアラボラトリー研究主務．主にコンテクストアウェア技術および情報抽出の研究開発に従事．

菊池匡晃
masaaki11.kikuchi@toshiba.co.jp
2006 年大阪大学大学院工学研究科知能・機能創成工学専攻修士課程修了．同年（株）東芝入社．現在，研究開発センター知識メディアラボラトリー主事．主にコンテクストアウェア技術および情報抽出の研究開発に従事．