

マイクロブログストリーム中の位置参照表現に着目した 実世界イベントの時空間可視化

伊藤 正彦^{†,††} 吉永 直樹^{†,††} 豊田 正史^{††}

[†] 情報通信研究機構ソーシャル ICT 推進研究センター 〒153-8505 東京都目黒区駒場 4-6-1

^{††} 東京大学生産技術研究所 〒153-8505 東京都目黒区駒場 4-6-1

E-mail: [†]{imash, ynaga, toyoda}@tkl.iis.u-tokyo.ac.jp

あらまし 本稿では、マイクロブログ記事の位置参照表現を利用し、投稿中の各単語の時空間的な局所性を単位領域ごとに算出し、これを実世界イベントとしてワードクラウド表現により地理空間中に可視化する手法を提案する。提案手法では、まず、マイクロブログストリーム中の各投稿中に現れる地名や施設名などの位置参照表現を認識する。次に、投稿中の各単語を、認識した地名や施設名の位置に紐付けて、局所的に観測された単語をローカルイベントとして認識する。地理空間ワードクラウドは時間経過に従って、そのサイズ、位置が変化し、時空間におけるイベントの影響を探索できる。実世界イベントとしては、限定された場所で行われるイベントに加えて、広範囲におよぶイベントなども考えられる。そこで、ローカルイベント群からグローバルイベントを抽出し、それらを 3 次元空間における多層地理空間ワードクラウド表現により可視化する。過去 5 年分からのツイッターデータを用いて提案システムを構築し、いくつかの応用事例を用いて提案システムの有用性を示す。

キーワード ソーシャルメディア, テキスト解析, 時空間情報可視化

1. はじめに

東京やニューヨークのような巨大都市では、様々な種類のイベント（大規模集客イベント、交通事故、自然災害など）が日々発生しており、発生する領域、期間および影響範囲は様々である。1 つのイベントの影響が広い領域に及んでいくこともあり、このような変化し続けるイベントを理解することは、都市計画、交通管理、防災・災害対応の観点から重要になってくる。

モバイルデバイスの普及に伴うソーシャルメディアの急激な発達により、過去には考えられなかった規模の人々の行動および考えを直接観察することが可能になっている。特に、Twitter などのマイクロブログは、場所と時間に関連する大量かつ詳細なイベント情報をリアルタイムに発信かつ取得可能にし、最近では、ほぼすべてのイベントにおいて、誰かしらがそこで何が起きているかをマイクロブログに投稿している。

このような投稿から時空間イベントを理解するために、地図上にワードクラウド表現でイベント情報を可視化する手法が提案されている [9], [14]。しかしながら、ユーザにより位置情報をタグ付けされた投稿は全体のうち 1% にも満たず、これらの手法では、本来ならイベントに関係するはずの大量の投稿を見落していることになる。また、イベントが地域限定のものか、広い領域に影響があるものなのかまで考慮して可視化をしておらず、それらが時間的・空間的に変化していく様子を視覚的に理解することも難しい。

本稿では、マイクロブログストリーム中の投稿から、投稿時間、ジオタグおよび位置参照表現を利用し、ローカルイベント群およびグローバルイベントを抽出し、それらを 3 次元多層ワードクラウド表現により地理空間中に可視化する。提案手法

では、まず、マイクロブログストリーム中の各投稿中に現れる地名や施設名などの位置参照表現を認識する。次に、投稿中の各単語を、認識した地名や施設名の位置に紐付けて、局所的に観測された単語をローカルイベントとして認識する。抽出されたローカルイベントは、地理空間ワードクラウドとして可視化される。地理空間ワードクラウドは時間経過に従って、そのサイズ、位置が変化し、時空間におけるイベントの影響を探索できる。多くの実世界イベントは、野球の試合やコンサートのような限定された場所で行われるが、地震やマラソンのような広範囲におよぶイベントなども考えられる。そこで、ローカルイベント群からグローバルイベントを抽出し、それらを 3 次元空間の別レイヤに可視化する。この多層地理空間ワードクラウドにより、個々のローカルイベントおよびグローバルイベントの探索、さらには、イベント間の関係の探索が可能になる。実際に、過去 5 年分からのツイートデータを用いて提案システムを構築し、いくつかの応用事例を用いて提案システムの有用性を示す。

2. 関連研究

2.1 マイクロブログの投稿への空間位置関連付け

既存のマイクロブログ投稿の時空間可視化研究の多くは、投稿に用いたモバイル端末の GPS 情報などを直接ユーザが付与した投稿のみを解析の対象としている [2], [9], [14] が、そのような投稿は例えば Twitter の場合、全投稿の 1% に過ぎないことが知られており [10] 多くは Foursquare のような位置情報サービスを介して発信されたもので、内容に乏しい。そのため、そのような投稿のみを分析の対象とした場合、大多数のローカルイベントは見逃すことになりかねない。ユーザの投稿位置を

自動推定しようという試みもあるが[7]、その精度は 30km 程度 [11] と細かい粒度で投稿の位置を推定することは難しく、本研究で取り扱うような局所性の高いローカルイベント検出に用いるにはまだ発展途上の段階にある。さらに、ユーザは必ずしも投稿位置に関係する情報を投稿するとは限らないという問題もある。

この問題に対し、Watanabe らは Foursquare の投稿から、位置的曖昧性の少ない施設名を取り出し、これを用いて投稿に位置情報を関連付けることを試みている [17]。我々はこの研究を参考に、施設名と共に起る投稿中の名詞にも位置情報を暫時的に伝搬することで大規模な位置情報参照表現辞書を構築し、これを用いて投稿の位置情報を推定することを試みる。

2.2 ワードクラウド表現によるイベント可視化

マイクロブログストリームから抽出したイベントを地図上にワードクラウド表現で可視化し探索可能にする研究の多くは、グローバルイベントのみに着目している [9], [14], [18]。これらに対して、我々の手法では、マイクロブログストリームから局所性の高いローカルイベントと、広範囲に影響がおよぶグローバルイベントを自動抽出し、それらを 3 次元空間の異なるレイヤ上にワードクラウド表現で可視化を行う。

ワードクラウドには様々なレイアウト手法が考案されている [8], [12], [15]。Cui および Wu らは、異なるタイムスタンプからなる複数のワードクラウド間で、同じ単語の位置変化を最小にするレイアウト手法を提案している [4], [16] が、彼らの手法は地理的な制約までは考慮していない。

2.3 2.5 次元表現を用いた多層情報可視化

3 次元空間に複数の可視化平面を重ねる 2.5 次元可視化手法としては様々なものが提案されており、関連する情報の可視化、時間変化の可視化、もしくは異なる可視化表現を用いた比較探索を可能にしている [1], [3], [6], [13]。VisLink [3] の手法では、ワードクラウド層と地図層との間にリンクを張ることによりそれらの間の関係性の探索を可能にするが、地理空間での単語出現位置を考慮した複数ワードクラウド表現およびそれらの時間変化までは考慮していない。

3. データセット

本研究ではマイクロブログデータとして、東京大学、喜連川・豊田研究室で 2011 年 3 月 11 日から Twitter API を用いて収集したツイートデータを用いている。ツイートの収集は、30 人の著名 Twitter ユーザを初期ユーザセットとして過去の投稿を取得するところから開始している。その後、彼らのタイムライン上でリツイートやメンションが行われたユーザを新たな対象としてユーザセットに加え、同様に過去の投稿を取得する。このように、ユーザセットの拡大とタイムラインのトラッキングを繰り返すことで、2015 年の時点で 200 万以上のユーザから約 250 億ツイートを蓄積している。

4. 提案手法

提案手法では、マイクロブログストリーム中の投稿から、ユーザの指定した時空間粒度でリアルタイムにイベントを検知し、

地図上にワードクラウドとして可視化し、ユーザの指定した時間間隔で更新する。以下で、マイクロブログストリームから一時的ローカルイベントを検出する手法と、検出したイベントを地図上に可視化する手法をそれぞれ説明する。

4.1 投稿からのローカルイベント検出

Twitter では、位置情報が付与された投稿は全体のわずかに 1% に過ぎないため、我々は投稿内容を元により多くの投稿を地図上の位置に関連づけることにした。具体的な手順としては、位置情報を含む地名辞書を用意して、地図が辞書に登録された地名を含んでいるとき、投稿をその地名の位置に関連付ける。ここで問題となるのは、地名は必ずしも地図上の一点を指すとは限らず、より広い範囲や、場合によっては複数の位置を指す場合があることである。そこで我々は、地名に限らず特定の位置と強い関連を持つ位置参照表現を含む位置参照表現辞書を構築し、これを用いて投稿を特定の位置に関連づけることにした。

提案手法では、まず Foursquare のような位置情報サービスを介して投稿されたツイートを、我々の保持するツイートアーカイブから列挙し、その投稿に含まれる施設名と位置を組み合わせて位置参照表現辞書を構築した。続いてこの辞書に、特定の位置と強く関連付けられる名詞を追加することで辞書の規模を拡大する。具体的には、投稿を形態素解析し、テキストから名詞を検出したのち、その名詞を投稿中で共起する（上記の辞書に登録されている）施設名の位置と関連付ける。一定頻度現れた名詞について、関連付けられた位置の分散が閾値を下回るとき、その名詞とその平均位置を位置参照表現辞書に追加する。このようにして得られた位置参照表現辞書のエントリ数は 38,504 であった。

続いてこの位置参照表現辞書を用いてマイクロブログストリームの投稿中の地名と施設名を認識する。もし投稿が登録された位置参照表現を含んでいたら、その投稿はその位置参照表現に紐づけられた位置に関連付ける。このようにして地図上の位置に関連付けられた投稿を、位置情報が直接付与されたツイートと区別して、関節的位置情報付きツイートと呼ぶ。

我々が収集した位置参照表現辞書はマイクロブログストリームで未来に発生するイベント名は含んでいないため、一時的なローカルイベントを位置情報付きツイートから以下の手順で検出する。我々はまず、形態素解析を用いて投稿中の名詞と未知語を検出する。検出した語句 w_i は、投稿の位置と関連付けられる。ユーザが指定した位置グリッド集合 $\mathcal{G}$ と時間間隔について、TF-IGF (term frequency-inverse grid frequency) と呼ぶスコアを計算し、特定の時空間グリッド $g_j \in \mathcal{G}$ に特有の語句を一時的ローカルイベントとして検出する：

$$\text{TF-IGF}(w_i, g_j) = \frac{\text{freq}(w_i)}{\sum_{w_i \in g_j} \text{freq}(w_i)} \log \frac{|\{g \in \mathcal{G}\}|}{|\{g \in \mathcal{G} : w_i \in g\}|}.$$

ここで、IGF を過去の時間間隔にまたがって計算することでイベントのバーストを検知する。各グリッドごとに TF-IGF スコアの top- n を一時的なローカルイベントとして可視化エンジンに入力する。

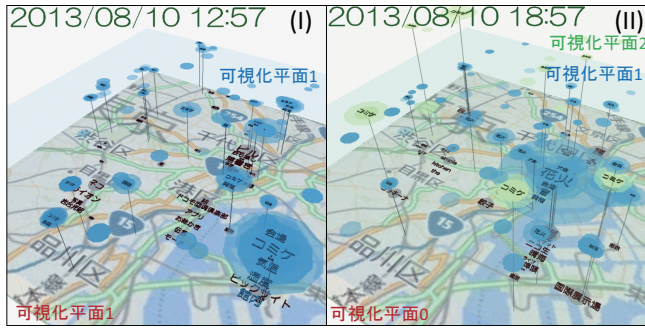


図 1 複数の大規模集客イベント発生時における多層地理空間ワードクラウド可視化事例 (2013 年 8 月 10 日、コミックマーケットと東京湾大華火祭が東京湾周辺で同日に行われた。)

4.2 多層地理空間ワードクラウド可視化

抽出された語句は地図上に多層ワードクラウドとして可視化される。我々の開発している可視化エンジンは、限定された場所で起こるローカルイベント群から広範囲に広がるグローバルイベント群を区別し抽出する。抽出されたグローバルイベントおよびローカルイベント群は、3次元空間における異なる可視化平面上にそれぞれワードクラウドとして可視化される。可視化平面は3次元空間の高さ方向に堆積され、3次元多層ワードクラウドを構成する(図1)。一つの可視化平面にすべてのイベントを重ねないように可視化すると、レイアウトの都合上実際の位置とは離れた場所にイベントを表示する必要が生じる。そこで、複数の可視化平面を用いることで各可視化平面内での場所のずれを最小限に抑えた表示が可能とする。これにより、異なるワードクラウド間でのイベントの位置関係の比較も容易になる。一方、これにより、重なり合った可視化平面間でのイベントが重なって表示される問題が生じるが、この問題への対応として、可視化平面の表示／非表示の選択、および堆積する高さの変更をインタラクティブに行える。また、3次元空間においてはユーザは視点を自由に変更できるため、重なり合いの少ない角度での探索を行える。イベントの時間変化は多層可視化平面内のワードクラウド群をアニメーションさせることにより可視化する。これにより、地図上に実際の雲が動いているような表現を実現している。

可視化エンジンでは、まず、ユーザが選択している表示領域内において可視化対象となる top- k イベントを決定する。そのために、表示領域内の各グリッドにおいて、それぞれの語句に関する時間減衰を考慮した TF-IGF スコアの合計を求め、top- k イベントを決定する。

グローバルイベントを抽出するために、ユーザが自由に設定可能な二種類の閾値を用意した：

- **クラスタ数:** ある語句が特定の位置(緯度および経度)においてのみ関連付いて投稿されているか、複数の位置に関連付いて投稿されているかを確認する。そのために、過去 T 期間におけるその語句の出現位置を密度ベースのクラスタリング手法である DBSCAN [5] によりクラスタリングを行う^(注1)。クラ

スタ数が閾値を超えた場合、その語句をグローバルイベントとして扱う。^(注2)

- **位置分散:** クラスタ数は少ないが広い領域にわたって出現している語句も存在しており、これらをグローバルイベントとして抽出するために、各語句の出現位置の分散を計算している。分散の値が閾値を超える場合、その語句をグローバルイベントとして扱う^(注3)。

Top-1 から top- k イベントまでスコア降順にグローバルイベントか否かを判断する。グローバルイベントと判断された場合、新たな可視化平面をローカルイベント群を表示する地表面の上に追加し可視化する。さらに、残りの上位イベント(を表す語句)が抽出されたグローバルイベント(を表す語句)と共に共起するかを確認する。共起する場合は、それらのイベントを抽出されたグローバルイベントの可視化平面に追加し、以降、それらのイベントに関してグローバルイベントかどうかの確認は行わない。

図1は、二つのグローバルイベントが発生した日における多層ワードクラウド可視化の例である。一つ目のイベントは東京ビッグサイトで行われるコミックマーケット(コミケ)で、50万人以上が日本中から集まる^(注4)。コミケの開催日には、さまざまな場所で連動イベントが行われたり、コミケの行き帰りの人々およびそれを目撃した人々が様々な場所でコミケに関連した投稿を行う。そのため、朝の早い段階からグローバルイベントとして抽出され、可視化平面1上に表示されている(図1(I))。また、真夏に大量の人がビッグサイトに集まったことによる気温上昇が話題になったことを表す関連イベントが可視化平面1上に追加されていることが確認できる。二つ目のイベントは晴海埠頭で行われた東京湾大華火祭で、夕方過ぎからグローバルイベントとして抽出されている。そのため、図1(II)では二つの可視化平面上にそれぞれのグローバルイベントが可視化されていることが分かる。また、地表面である可視化平面0上には、ローカルイベントである“ビル”“落書き”“取り壊し”などが表示されていることが確認できる(図1(I)および(II))。

各イベントの規模や位置、特に、マラソンや竜巻のような移動するイベントの規模や位置を追跡するため、ワードクラウド表現を応用しそれらの可視化を行う。各語句クラスタは一つの円として表現される。そして、その語句が円の中心に表示される。クラスタの中心は、スコアと時間減衰を考慮した出現位置の加重平均により求める。円と語句の表示サイズはクラスタに含まれる出現の合計スコアの平方根で定義する。また、語句出現のばらつき具合を観測するために、語句の各出現も小さな円として表示している。各円のサイズはスコアを表し、透明度は時間減衰を表す。

複数のイベントがほぼ同じ場所に出現した場合、それらが重なり合うことを防ぐ必要がある。今回、スコアの小さな語句をスコアの大きな語句の下方もしくは上方に移動することで重なり

^(注1) MinPts) はユーザが自由に設定可能である。

^(注2) : 本稿の事例ではクラスタ数閾値は 6 に設定している。

^(注3) : 本稿の事例では位置分散閾値は 0.006 に設定している。

^(注4) : <http://www.comiket.co.jp/archives/Chronology.html>

^(注1) : DBSCAN における二つのパラメータ (距離 ϵ および 最小ポイント数

りを防止する、という単純な方法を用いている。ほとんどの語句においては、語句の幅よりも高さの方が小さいため、左右よりも上下に移動する方が移動量が小さくすむ場合が多く、イベント間の相対位置関係を維持できるため、このような手法を用いた。

5. 探索事例

5.1 東京マラソン 2012

図 2 は東京マラソン 2012^(注5) が開催された 2012 年 2 月 26 日におけるイベント変遷を可視化した例である。

東京マラソンでは毎年約 35,000 のランナーが参加しており 100 万人以上の人々が沿道から観戦をしている。東京マラソンのコース^(注6) は東京の西から東への移動、さらに南北間での 2 度の折り返しを含む (図 2 (II) における $A \rightarrow B \rightarrow C \rightarrow B \rightarrow D \rightarrow B \rightarrow E$)。完走する多くのランナーは約 3 時間から 6 時間でゴールしており、各主要観戦ポイントにおける通過時間は表 1 のようになっている。また、多くの人々がランナーを応援するために主要観戦ポイント間を移動し、状況をツイッターなどに投稿している。

図 2 (I) から (VI) では、“東京マラソン”という大きなグローバルイベントとそれに関連する様々なイベントとそれらの時間変遷を観測できる。それぞれの時間のスクリーンショットにおけるイベントのサイズと場所は、その時間におけるランナーの走行位置を反映している。関連イベントのサイズと場所からイベントの詳細状況を探索することが可能である。

4.1 節の手法により取得した間接的位置情報付きツイート数は 11,927 になり、ユーザにより直接に位置情報が付与されたツイート数 2,338 を大きく上回る。

5.2 有楽町火災

図 3 では、2014 年 1 月 3 日における JR 有楽町駅の沿線火災発生後のイベント時間変遷を可視化している。火災は午前 6 時半ごろに発生し、有楽町駅周辺が煙に包まれた影響で、東海道新幹線、JR 山手線などが長時間にわたり運転を見合わせた。1 月 3 日はまだ正月休みであり、多くの人々が寺社に初詣に訪れる期間であることもあり、多くの人々の移動に影響を及ぼした。

図 3 (I) においては、まだ早朝で人々の移動が始まっていないこともあり、“火災”イベントはローカルイベントの一つである。図 3 (II) においては、多くの人々が移動を開始し各所で移動できないトラブルに巻き込まれたため、“火災”イベントはグローバルイベントとして検出されていることがわかる。さらに、図 3 (III) の可視化平面 2 上には、主要な寺社周辺に“初詣”イベントが出現していることが確認できる。

4.1 節の手法により取得した間接的位置情報付きツイート数は 22,482 になり、ユーザにより直接に位置情報が付与されたツイート数 3,070 を大きく上回る。

6. まとめ

本稿では、マイクロブログストリーム中の投稿に対し、位置参照表現等を利用することで空間位置に関連付けし、ローカルイベントおよびグローバルイベントを抽出し、多層地理空間ワードクラウドによりその時間変遷を探索するための可視化手法を提案した。また、東京マラソンと有楽町駅火災という東京都内で発生した二つの大規模イベントを事例として時空間イベントの探索を行い、提案手法の有用性を示した。

文 献

- [1] Brooks, S., and Whalley, J. L. Multilayer Hybrid Visualizations to Support 3D GIS. *Computers, Environment and Urban Systems* 32, 4 (2008), 278–292.
- [2] Chae, J., Thom, D., Bosch, H., Jang, Y., Maciejewski, R., Ebert, D. S., and Ertl, T. Spatiotemporal Social Media Analytics for Abnormal Event Detection and Examination using Seasonal-trend Decomposition. In *VAST* (2012), 143–152.
- [3] Collins, C., and Carpendale, S. VisLink: revealing relationships amongst visualizations. *IEEE Trans. Vis. Comput. Graph.* 13, 6 (2007), 1192–1199.
- [4] Cui, W., Wu, Y., Liu, S., Wei, F., Zhou, M. X., and Qu, H. Context Preserving Dynamic Word Cloud Visualization. In *PacificVis* (2010), 121–128.
- [5] Ester, M., Kriegel, H., Sander, J., and Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *KDD* (1996), 226–231.
- [6] Itoh, M., Yoshinaga, N., Toyoda, M., and Kitsuregawa, M. Analysis and Visualization of Temporal Changes in Bloggers’ Activities and Interests. In *PacificVis* (2012), 57–64.
- [7] Kinsella, S., Murdock, V., and O’Hare, N. “i’m eating a sandwich in glasgow”: Modeling locations with tweets. In *SMUC* (2011), 61–68.
- [8] Koh, K., Lee, B., Kim, B. H., and Seo, J. ManiWordle: Providing Flexible Control over Wordle. *IEEE Trans. Vis. Comput. Graph.* 16, 6 (2010), 1190–1197.
- [9] MacEachren, A. M., Jaiswal, A. R., Robinson, A. C., Pezanowski, S., Savelyev, A., Mitra, P., Zhang, X., and Blanford, J. SensePlace2: GeoTwitter Analytics Support for Situational Awareness. In *VAST* (2011), 181–190.
- [10] Middleton, S., and Middleton, L. Real-time crisis mapping of natural disasters using social media. *IEEE Intelligent Systems* 29, 2 (2014), 9–17.
- [11] Schulz, A., Hadjakos, A., Paulheim, H., Nachtwey, J., and Mühlhäuser, M. A multi-indicator approach for geolocalization of tweets. In *ICWSM* (2013), 573–582.
- [12] Seifert, C., Kump, B., Kienreich, W., Granitzer, G., and Granitzer, M. On the Beauty and Usability of Tag Clouds. In *IV* (2008), 17–25.
- [13] Shen, Z., Ogawa, M., Teoh, S. T., and Ma, K.-L. BiblioViz: a System for Visualizing Bibliography Information. In *APVis* (2006), 93–102.
- [14] Thom, D., Bosch, H., Koch, S., Wörner, M., and Ertl, T. Spatiotemporal Anomaly Detection through Visual Analysis of Geolocated Twitter Messages. In *PacificVis* (2012), 41–48.
- [15] Viégas, F. B., Wattenberg, M., and Feinberg, J. Participatory Visualization with Wordle. *IEEE Trans. Vis. Comput. Graph.* 15, 6 (2009), 1137–1144.
- [16] Wu, Y., Provan, T., Wei, F., Liu, S., and Ma, K. Semantic-Preserving Word Clouds by Seam Carving. *Comput. Graph. Forum* 30, 3 (2011), 741–750.

(注5) : <http://www.marathon.tokyo/info/past/2012/>

(注6) : <http://www.marathon.tokyo/info/course/>

表 1 東京マラソン 2012 におけるランナーのおおよその通過時刻

	A	→ B	→ C	→ B	→ D	→ B	→ E
トップランナー	9:10	9:40	9:55	10:13	10:34	10:55	11:17
3 時間ランナー	9:11	9:53	10:15	10:41	11:10	11:40	12:11
6 時間ランナー	9:24	10:49	11:32	12:24	13:22	14:22	15:24

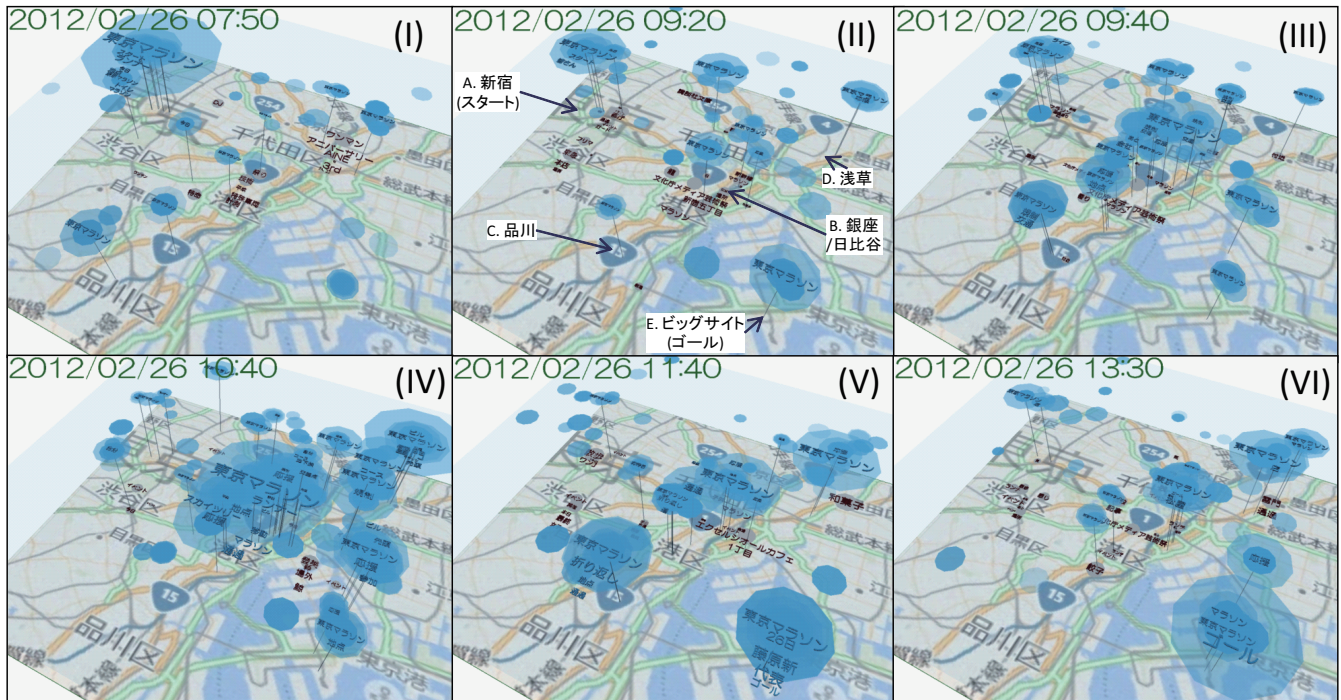


図 2 多層地理空間ワードクラウドによる東京マラソン 2012 開催日におけるイベント変遷の可視化

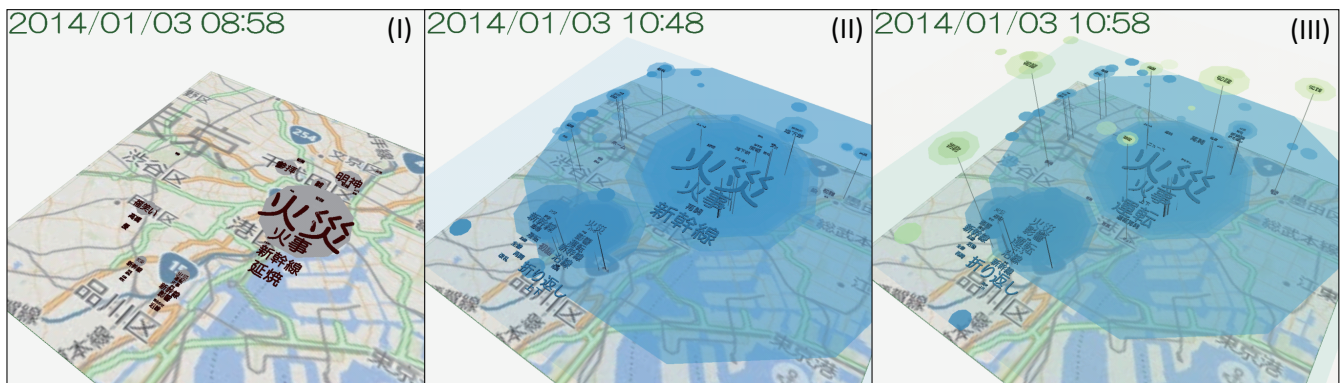


図 3 多層地理空間ワードクラウドによる, 2014 年 1 月 3 日有楽町駅における火災発生日におけるイベント変遷の可視化例

- [17] Watanabe, K., Masanao, M., Okabe, M., and Onai, R. Jasmine: A Real-time Local-event Detection System based on Geolocation Information Propagated to Microblogs. In *CIKM* (2011), 1091–1102.
- [18] Bosch, H., Thom, D., Heimerl, F., Puttmann, E., Koch, S., Kruger, R., Worner, M., Ertl, T. ScatterBlogs2: Real-time Monitoring of Microblog Messages through User-guided Filtering. *IEEE Trans. Vis. Comput. Graph.* 19, 12 (2013), 2022–2031.