Twitterからのジオタグ画像収集による視覚的イベント検出

金子 昂夢、柳井 啓司
電気通信大学 大学院情報理工学研究科
総合情報学専攻

背景

- スマートフォンの普及
  -位置情報付き画像
- Twitter
  -リアルタイムな投稿

位置情報付き画像ツイートの増加

目的

- Twitterストリームからイベントの検出
  - -天災、自然現象
  - -祭り、スポーツの試合
- 視覚的な検出
  - -代表画像の選択
  - -地図上に表示

画像と共に地図上に配置

関連研究: Twitterイベントマイニング

- 多くがテキスト情報のみを利用
- 榊らの研究 [WWW 2010]
  -ユーザをソーシャルセンサと捉えたモデルを構築
  -地震の発生を位置と共に高速に検出
- Leeらの研究 [ACM SIGSPATIAL WS 2010]
  -対象地域をより小さな領域に分割
  -ツイートの頻度をモニタリング

関連研究: Twitter画像マイニング

- Twitterに投稿された画像を利用
- 柳井の研究 [ACM ICMR 2012]
  -リアルタイムに投稿された画像をモニタリング
  -地図上に画像をマッピング
- 中地らの研究 [ICME WS 2012]
  -キーワードに関連する代表画像を抽出
  -位置や時間による違いを比較

関連研究: MediaEval SEDタスク

- Flickrデータを用いたSocial Event Detection
- Brennerらの研究[ACM ICMR 2012]
  -外部リソースを利用し高精度化
- Daoらの研究[ACM ICMR 2013]
  -ユーザ視点のデータベースを構築
  -類似画像検索で高精度化

システムの概要

1. イベントキーワードの検出

↓

2. キーワードの統合・補完

↓

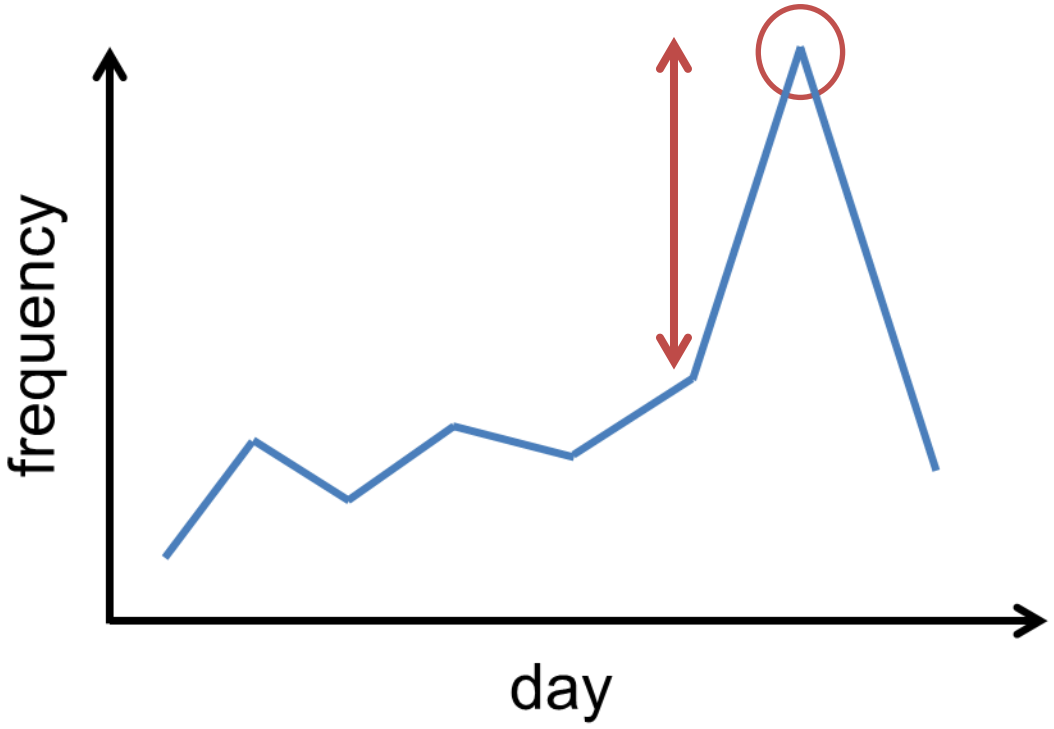
3. 画像のクラスタリング

↓

4. 画像と共に地図上に表示

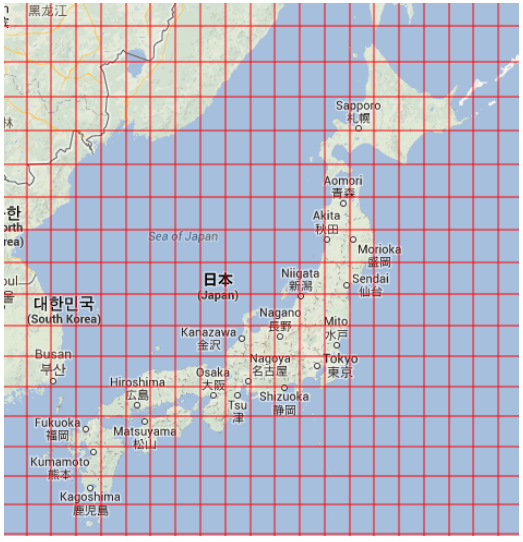
イベントキーワードの検出

- 1日毎に単語の頻度をカウント

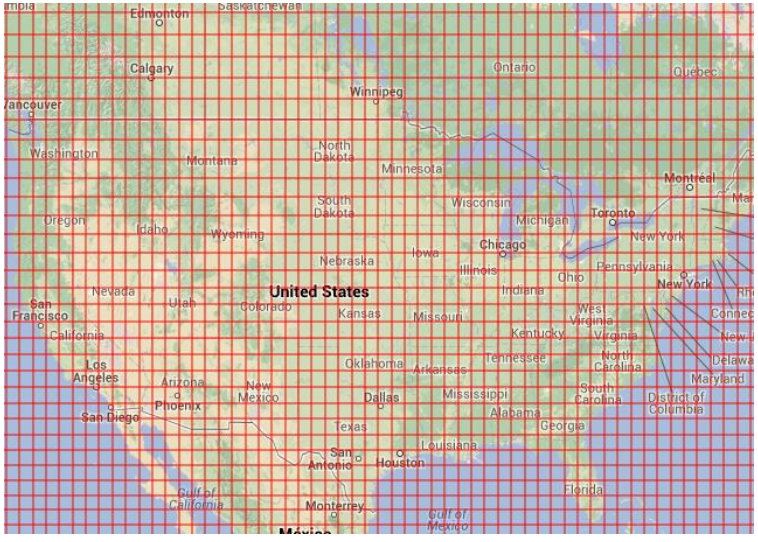


イベントキーワードの検出

- 形態素解析
  - -MeCab(http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html)
  - -TweetNLP(http://www.ark.cs.cmu.edu/TweetNLP/)
- 対象地域をより小さな領域に分割
  - -緯度・経度1度ごとのグリッドにより分割



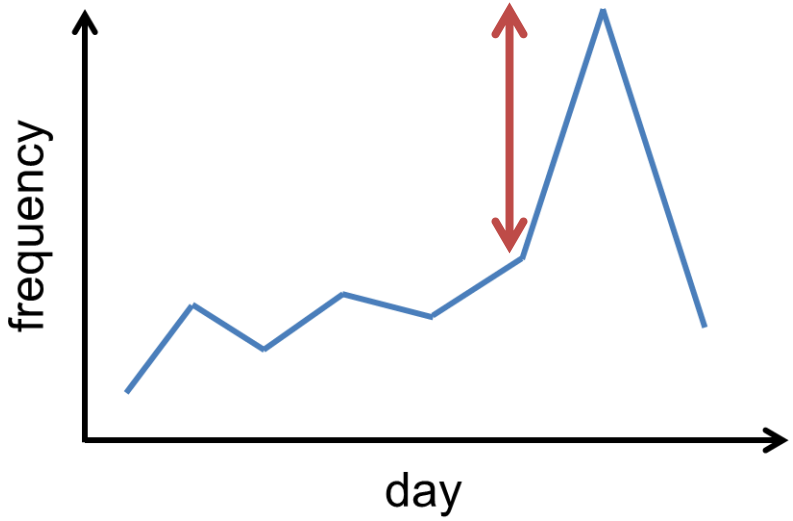
アメリカ



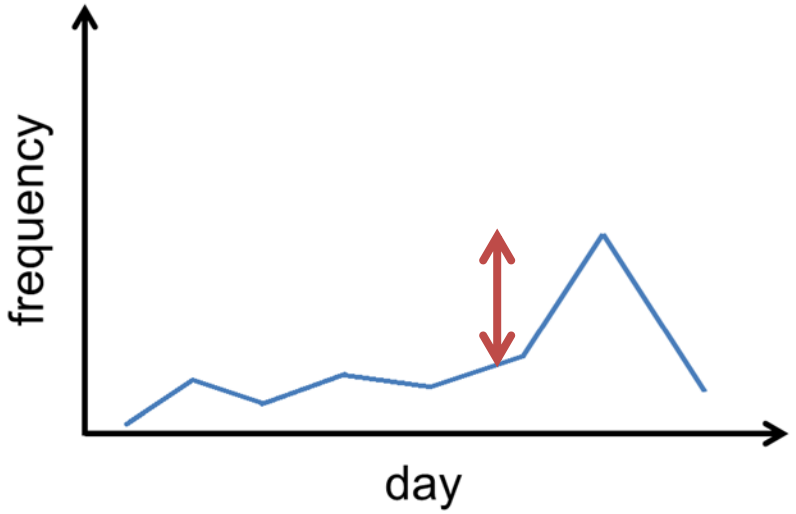
日本

イベントキーワードの検出

- ユーザ人口の違い
  -出現頻度上昇のスケールの違い



人口の多い地域



人口の少ない地域

イベントキーワードの検出

- 地域の重みを計算

$$W_{area} = \frac{\#users_{max} + sd}{\#users_{area} + sd}$$

- キーワードのスコアを計算

-50以上(日本), 200以上(アメリカ)

$$S_{kw,date,area} = (N_{kw,date,area} - N_{kw,date-1,area})W_{area}$$

キーワードの統合・補完

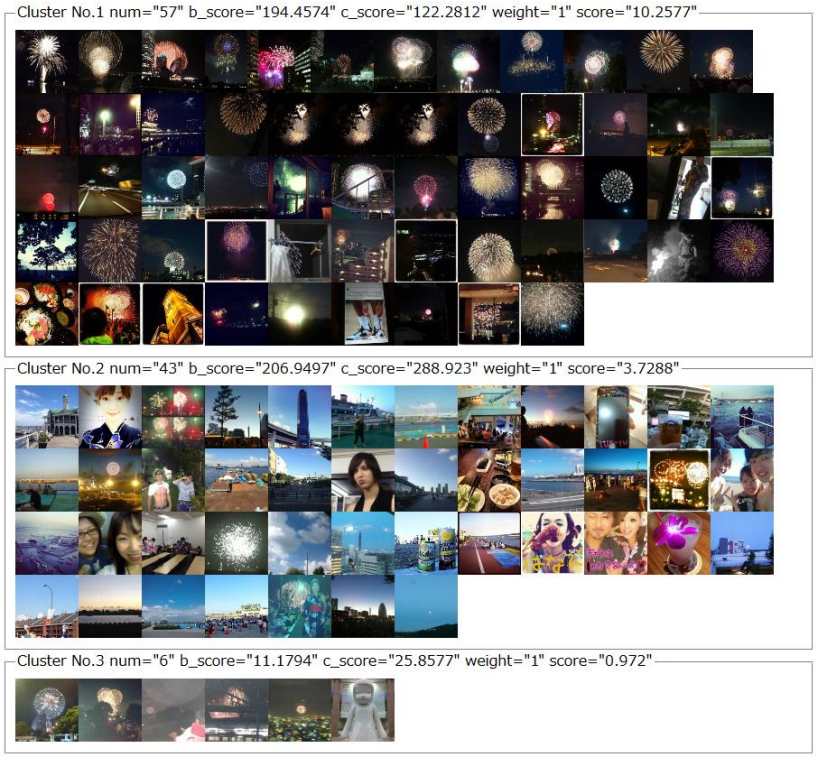
- キーワードの統合
  - -ツイートが50%以上同じ
  - -最もスコアの高いキーワードを利用

    「花火」、「大会」⇒「花火」
- キーワードの補完
  - -文字が80%以上同じ
  - -前後の文字を再帰的に補完

    「スカイ」⇒「スカイツリー」

画像のクラスタリング

- 画像特徴量
  - -SURFのBag-of-Features
  - -カラーヒストグラム
- Ward法
  - -階層型クラスタリング手法
  - -閾値300で終了

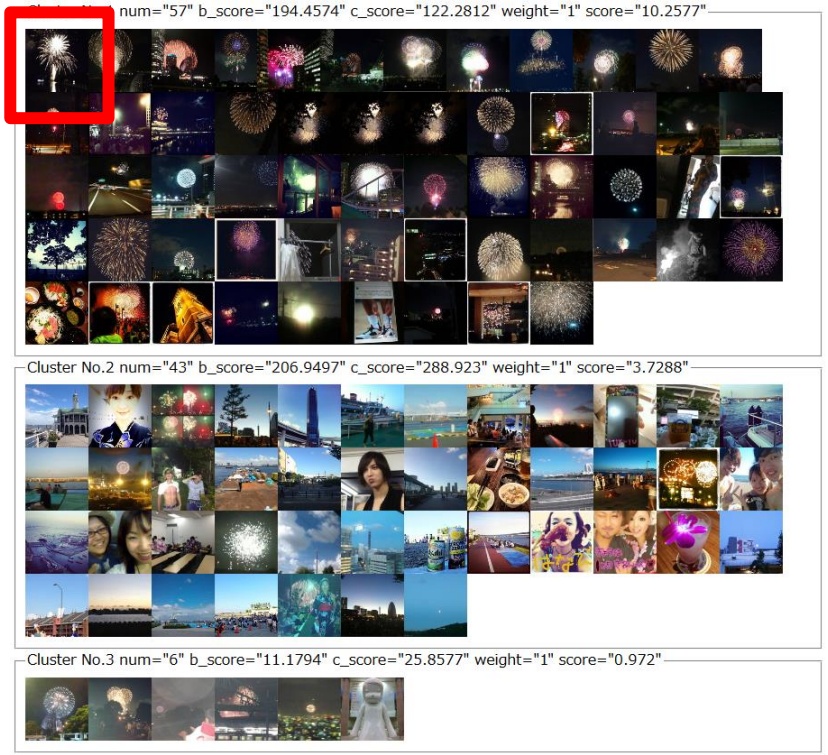


$$E(C) = \sum_{x \in C}((x_{BoF} - \overline{x_{BoF}})^2 w_{BoF} + (x_{RGB} - \overline{x_{RGB}})^2 w_{RGB})$$

代表画像の選択

- 代表クラスタの選択
  -クラスタのスコアを計算

$$V_C = \frac{\#photos_C{}^2}{E(C)} W_{area}$$



- 代表画像の選択
  -最大スコアのクラスタ中で
  中心に最も近いもの

- 低いスコアを持つクラスタを除外
  -5(日本)、20(アメリカ)以下

実験

- データセット1
  - -日本
  - -2011年2月10日から2012年9月30日
  - -約300万件の位置情報付き画像ツイート
- データセット2
  - -アメリカ
  - -2012年1月1日から2012年12月31日
  - -約1,700万件の位置情報付き画像ツイート

キーワードの検出結果

Keyword	Date
雪	11/02/2011
地震	11/03/2011
花火	06/08/2011
台風	21/09/2011
富士山	24/09/2011
アップル	06/10/2011
月食	10/12/2011
リエ	10/12/2011
クリスマス	24/12/2011
大晦日	31/12/2011
日の出	01/01/2012
ホタル	06/05/2012

日本

Keyword	Date
snow	09/01/2012
sunset	13/01/2012
Grammy	12/02/2012
Valentines	14/02/2012
SXSW	09/03/2012
Easter	08/04/2012
shuttle	17/04/2012
WWDC	10/06/2012
hurricane	26/08/2012
rainbow	05/09/2012
49ers	18/10/2012
NYE	31/12/2012

アメリカ

キーワードの統合・補完結果

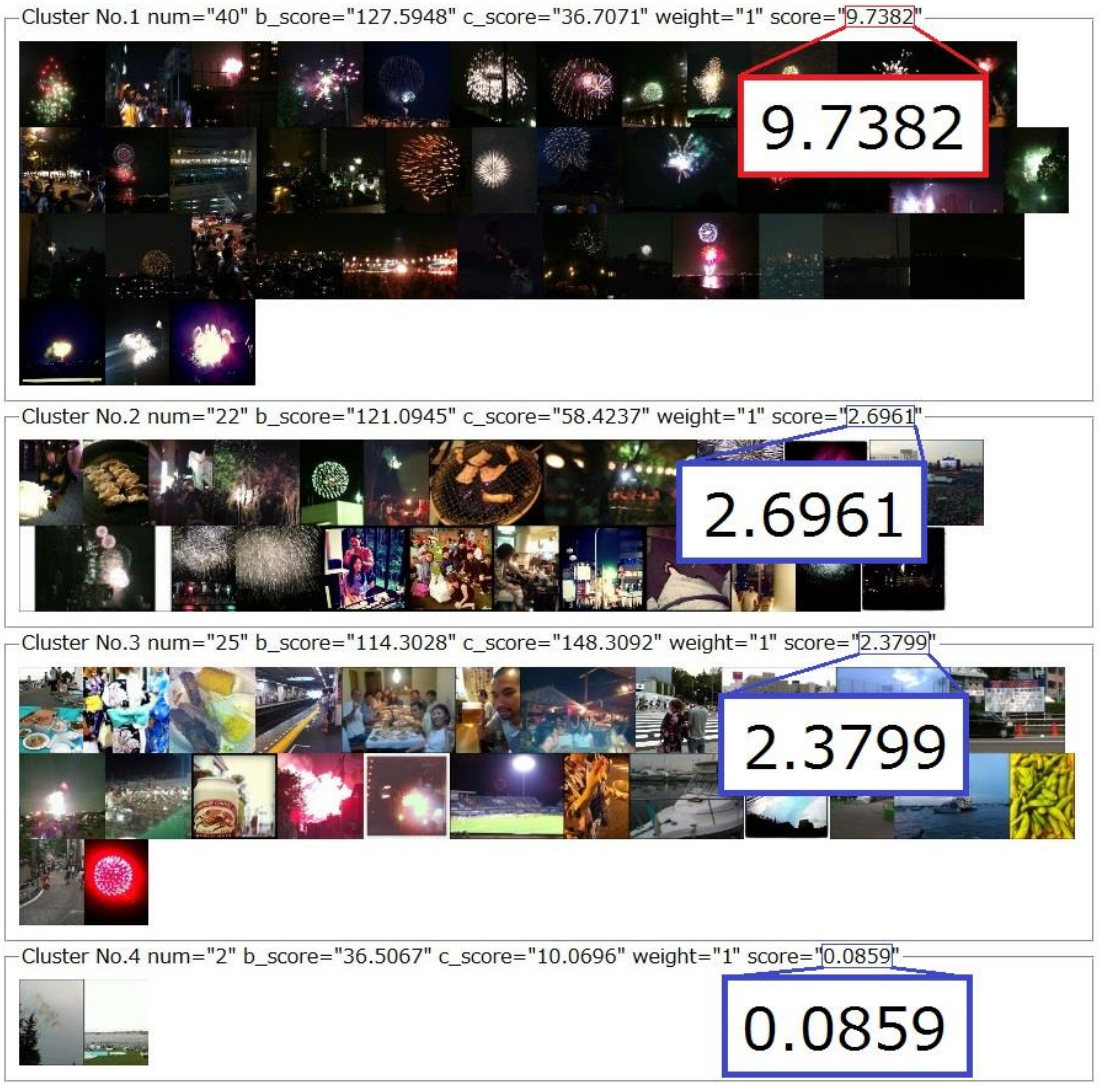
日本

統合・補完前	統合結果	補完結果
スカイ、ツリー	スカイ	スカイツリー
花火、大会	花火	花火
日食、皆既	日食	皆既日食
マラソン	マラソン	東京マラソン
リエ	リエ	ルミナリエ

アメリカ

統合・補完前	統合結果	補完結果
Rangers, Ballpark	Rangers	Rangers Ballpark
West, WWDC, Apple	WWDC	WWDC
Golden, Gate, Bridge	Golden	Golden Gate Bridge
Square, Times	Times	Times Square
Carnival, Electric, Daisy	Electric	Electric Daisy Carnival

「花火」のクラスタリング結果



「桜」のクラスタリング結果

Cluster No.1 num="32" b_score="89.4698" c_score="127.6658" weight="1.9642" score="9.2631"

Cluster No.2 num="24" b_score="77.7001" c_score="90.9009" weight="1.9642" score="6.7104"

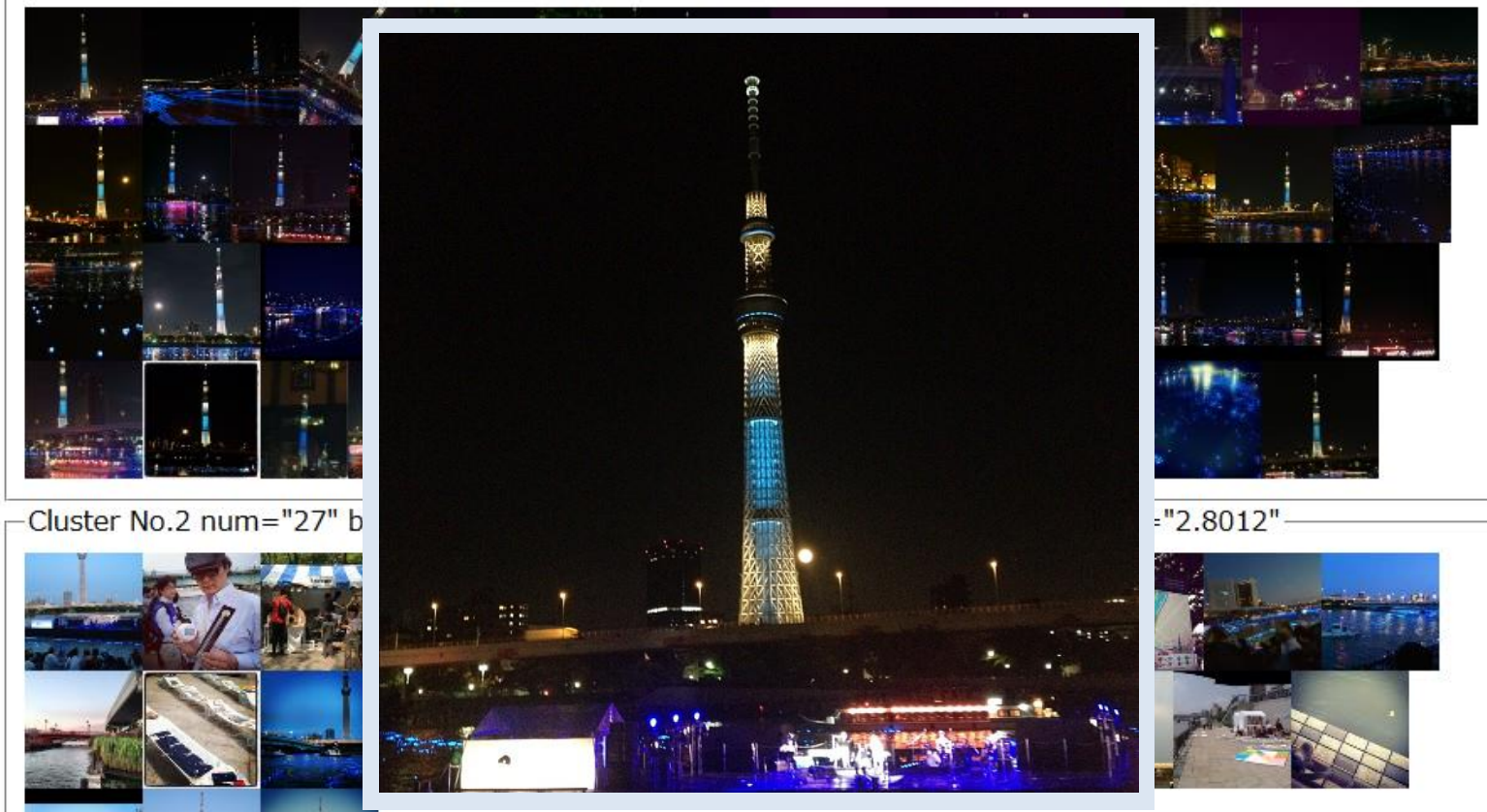
Cluster No.3 num="1" b_score="0" c_score="0" weight="1.9642" score="0.0002"

「Stanford Stadium」の結果

代表画像の選択

東京ホタル 2012/5/6

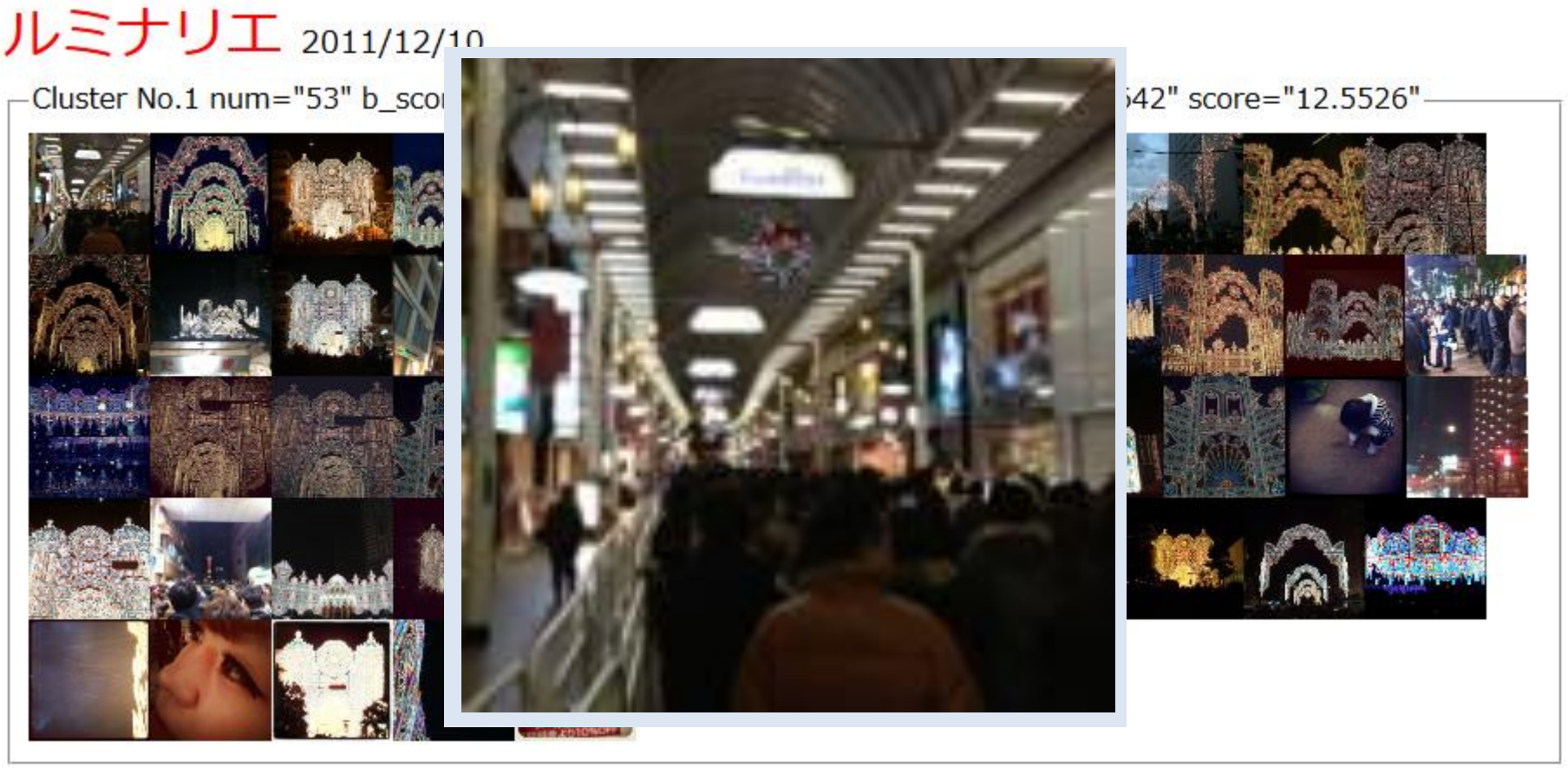
Cluster No.1 num="48" b_score="164.1649" c_score="44.3774" weight="1" score="11.0481"



Cluster No.2 num="27" b

"2.8012"

代表画像の選択



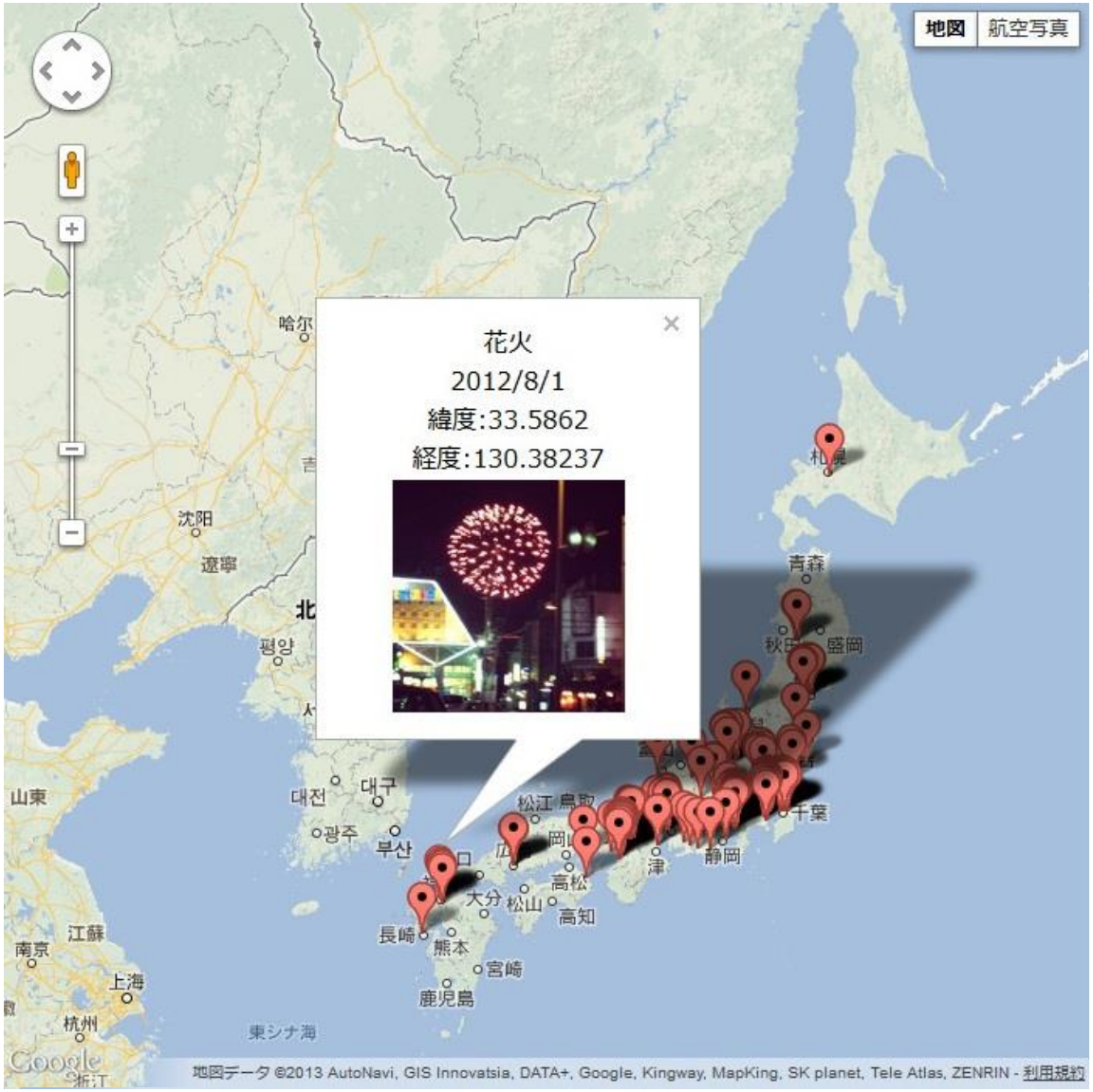
イベントの検出結果

- 検出されたイベントを地図上で表示
  -中心座標を計算
  -イベントの情報と代表画像を関連付け
- 検出結果のまとめ

	日本	アメリカ
イベント数	258	1676
代表画像の適合率	65.5%	72.5%

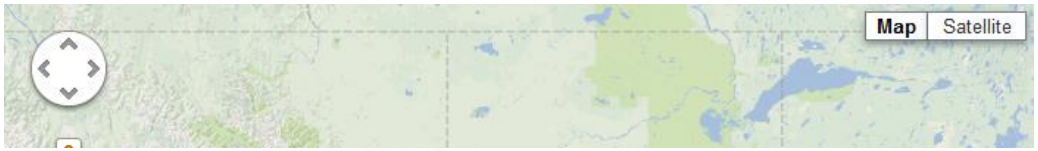
処理時間

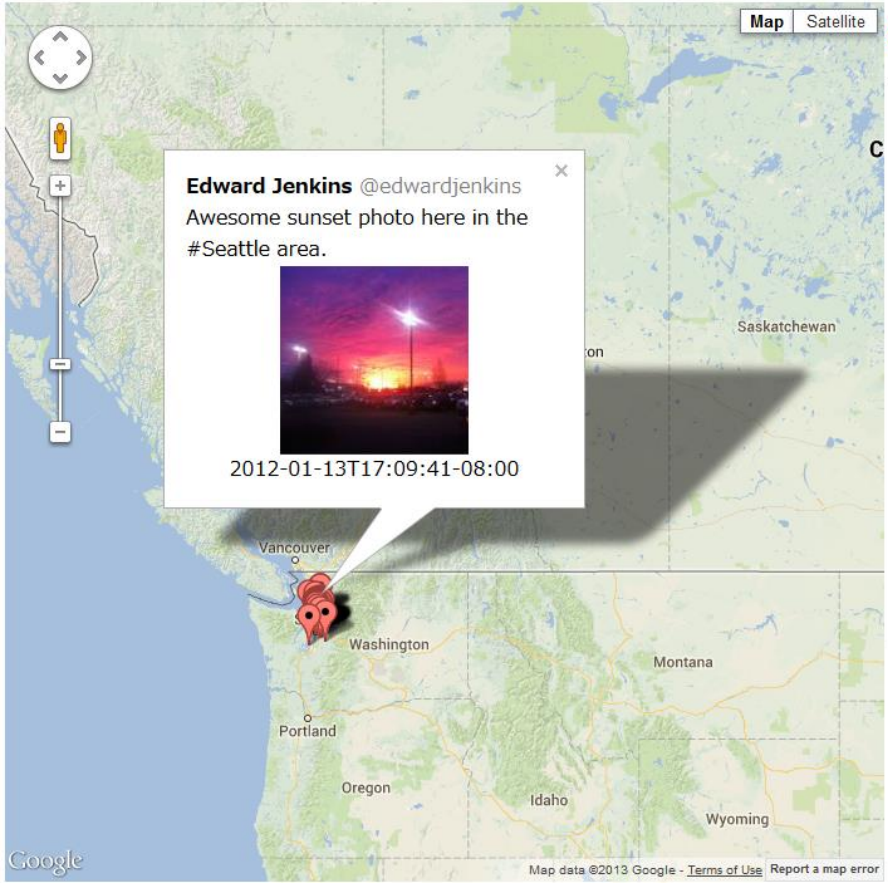
	日本	アメリカ
ツイート件数	約3,000,000,000件	約17,000,000,000件
形態素解析	20 min (1 core)	3 hours (1 core)
キーワードの検出	1 day (80 core)	3 days (80 core)
画像収集	10 days (全て)	1 days (161,389枚)
クラスタリング	1 hours (21,338枚)	7 hours (161,389枚)

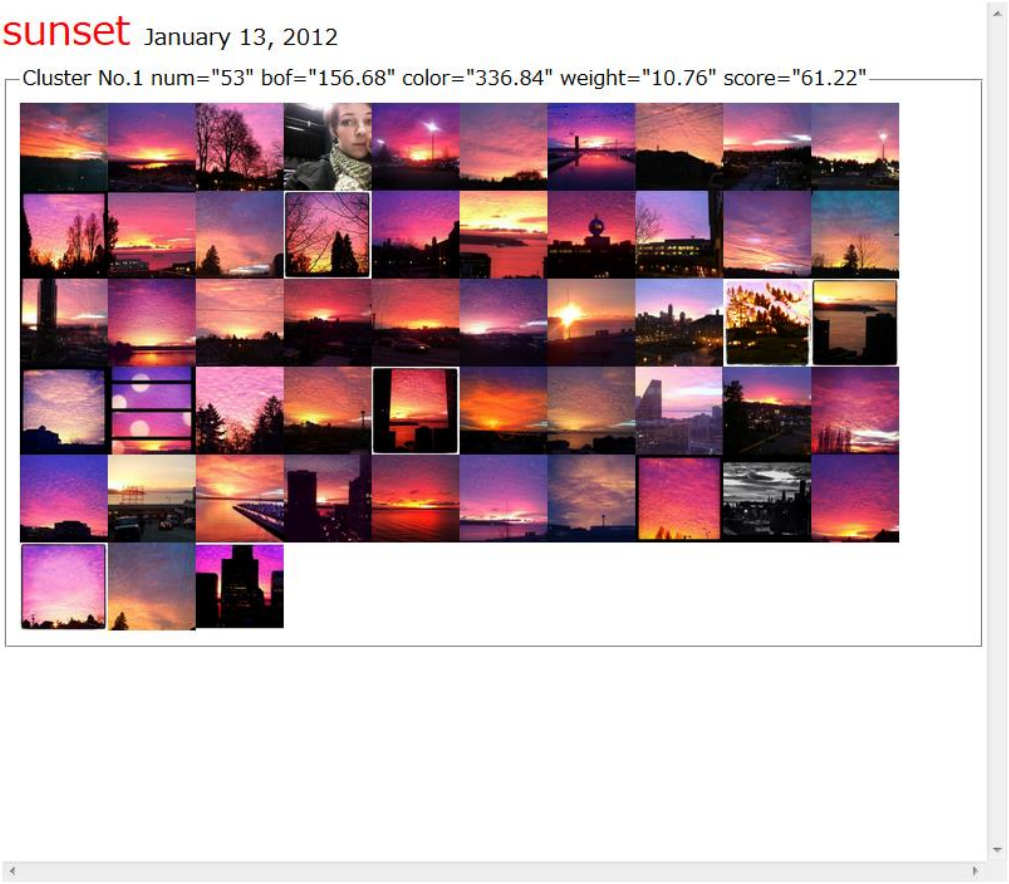


日本データ処理の結果

「River's Edge Music Festival」







「sunset」

まとめ

- ジオタグ画像ツイートを用いたイベント検出
  - -キーワードの検出
  - -代表画像の選択
- 検出結果
  - -日本データセットでは258件
  - -アメリカデータセットでは1676件
- 代表画像の適合率
  - -日本データセットでは65.5%
  - -アメリカデータセットでは72.5%

今後の予定

- より柔軟な検出
  -可変グリッドサイズ
  -イベントの期間の推定
- リアルタイムな検出
- 代表画像の選択手法の改良