1E4-5in2

類似部分時系列のクラスタ化に基づく 時空間データからのイベント抽出とその再起性・共起性の評価

Extraction of event clusters from spatio-temporal data based on clustering of similar time-series subsequence and evaluation of their recurrence and co-occurrence in the spatio-temporal neighborhood

森 啓太[*1] Keita Mori　森田博次[*2] Hirotsugu Morita　本田 理恵[*3] Rie Honda

[*1*2*3] 高知大学 Kochi University

A method to extract useful recurrence and co-occurrence of events in the spatio-temporal data is proposed. The spatio-temporal data is divided into time-series subsequences that are labeled by using SOM clustering. The labels are classified into "events" and "non-events" and spatio-temporal points that has the same "event" label and connected each other in the spatio-temporal neighborhood are defined as "event cluster". Recurrence and co-occurrence of "event clusters" are evaluated by using criterion based on support and confidence in association rules by dividing spatio-temporal field into sub-blocks and counting the event clusters by using scanning windows by using a-prior algorithm. The method is applied to the meteorological satellite images and the result is visualized as hot spots in the spatio-temporal field along with the extracted rules.

1. はじめに

セキュリティカメラによるモニタリングや数値シミュレーション，リモートセンシングによる地球観測等の様々な分野で大量の時系列画像や時空間データが取得されるようになっている．これらの大量の時空間データから時空間の変動パターンを抽出する事が出来れば，現象の理解や予測に役立てる事ができる．

時空間の変動パターンについては，DTW(Dynamic Time Warping)による類似性の解析，イベント抽出，相関係数による相関分析，統計学的なモデリングなどが検討されてきた．最近では時間，空間変動を統一的に非線形テンソル解析で扱おうとする手法[Matsubara 2014]が注目されている．

我々は，時間，空間方向に均一かつ密にデータが存在する時系列画像をターゲットとして，注目すべき時系列が与えられたときに，その相関係数からイベント，ならびにその時空間近傍の塊としてイベントクラスタを求め，イベントクラスタ毎の再起性を相関ルールで用いられる支持度，確信度をもとめ，その時空間分布を可視化する手法を提案した[Honda 2015]．しかし，この手法ではあらかじめターゲットの時系列を与える必要があった．今回はより多様な時系列変動を時空間データの集合体から自律的に発見したうえで，その再帰性，共起性を定量的に求め，その時空間分布を可視化してユーザーに提示する手法について検討する．

2. 手法

図 1 に本研究で行う時空間データの分析手法のイメージ図を示す．時空間データ内には様々な特徴的な時系列としてのイベントが存在するものと考えられる．ここでは時空間データを部分時系列に分割し，その特徴をクラスタリングすることで部分時系列のラベリングを行う．その際イベントと非イベントの分別も含むものとする．その後，類似した特徴を持つイベントの時空間の固まりをイベントクラスタとしてまとめ，類似した特徴を持つイベントクラスタの再起や，イベントクラスタが持つ特徴間の共起を，時空間の小領域ごとに評価するものとする．イベントクラスタの抽出とその評価についてはサブセクションに順に述べる．

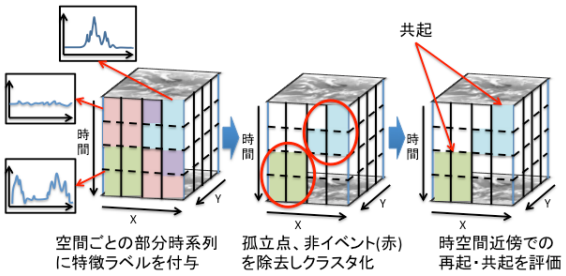


図 1. 共起性・共起性評価手法の概念図

2.1 イベントクラスタ抽出

(1) 部分時系列のクラスタリングとイベント抽出

時空間データの観測量(時系列画像であれば輝度値)を$I(x,y,t)$で表すものとする．ここで(x,y)は座標，tは時刻である．まず時空間データは長さωの重なりのない部分時系列に分割して

$$d(x,y,i)=\{I(x,y,\omega i),\cdots,I(x,y,\omega i+\omega-1)\}$$
$$D=\{d(x,y,i)|\,x=0,1,\cdots X-1,y=0,1,\cdots Y-1,\ i=0,1,\cdots T/n-1\}$$

で表す．ここでX,Y,Tは時空間領域のサイズを表すものとする．

時空間データ内のイベントの特徴の要約とラベリングには自己組織化マップ[Kohonen 2000]を用いる．自己組織化マップは入力層と競合層の2層のネットワークからなる教師なし学習アルゴリズムである．入力層にはl次元の入力データ$\{p|p_i\in R^l, i=1\ldots N\}$が配置されており，競合層には参照ベクトル$\{M_i|M_i\in R^l, i=1\ldots O\}$が2次元的に配置されている．自己組織化マップの学習では入力データとの類似度が最も高い勝者ベクトルM_c

連絡先：森啓太，高知大学大学総合人間自然科学研究科，高知市曙町 2-5-1，keitmori@is.kochi-u.ac.jp

を競合層から一つ決定し，M_cとその周りの参照ベクトルをp_iに近づけることで入力層の特徴を競合層に学習させる．

ここでは時空間の全点における長さωの部分時系列$d(x,y,i)$を要素とする時系列データ集合Dを入力として自己組織化マップによる学習を行うことで時空間データ内に存在する部分時系列の特徴を競合層に要約させる．学習の後，部分時系列と勝者ベクトルとの対応関係を用いて部分時系列に対してクラスタリングを行い，特徴ラベル$L(x,y,i)$を付与する．これにより部分時系列の特徴ラベルの時空間分布を得ることができる．

また，部分時系列には変動が小さくイベントとして意味のない部分時系列も多く含まれている．よって競合層の参照ベクトルの内，変動の小さいラベルを非イベントラベルとし，これを除いた部分時系列のみをイベントとして抽出して，以降の解析の対象とする．

(2) イベントクラスタの定義と抽出

時空間において隣接する同ラベルのイベントは同一の現象を示していると考えられるため，時空間内で塊状に存在するイベントをクラスタ化し”イベントクラスタ”として定義する．クラスタ化には先行研究[Honda 2015]同様，三次元データのラベリング処理[何 2009]をラベルタイプ毎に実施し，イベントクラスタのリストを取得する．

2.2 時空間の小領域ごとの再起性・共起性の評価

イベントクラスタの再起性・共起性は時空間の領域ごとに異なると考えられるため，時空間を小領域(ブロック)に分割し，そのブロックごとに再起性・共起性を調査する．イベントクラスタの再起性・共起性の評価は相関ルール[Agrawal 1994]における支持度と確信度を用いて行う．相関ルールではアイテム集合であるトランザクションの集合を調査し，同一のトランザクションに現れやすいアイテムの組み合わせをルールとして抽出する．そのため，時空間データ内でのイベントの再起・共起を相関ルールで表すためには時空間データにおけるトランザクションをどのように取得するかを定義する必要がある．以降，ブロック毎に行うトランザクションの取得方法について説明する．

(1) 時空間データにおけるトランザクションの定義

ブロックごとのトランザクション定義方法の概要図を図 2 に示す．分割したブロック(図 2 の黒太線枠)内に，時空間データにおける走査窓として時空間の窓(図 2 の赤枠)を定義する．この窓でブロック内を走査し，ブロック内に存在するイベントクラスタの数をその特徴ラベルごとにカウントしたものを“トランザクション”として定義する．イベントクラスタは広がりをもっているため，その一部のみが走査窓に含まれる場合がある．この場合は含まれる部分のそのクラスタ全体に対する体積比率が 25%を超えたときに，トランザクションに含まれるとしてカウントする．ここで走査窓のサイズが，抽出されるルールの時空間近傍を規定することに注意が必要である．

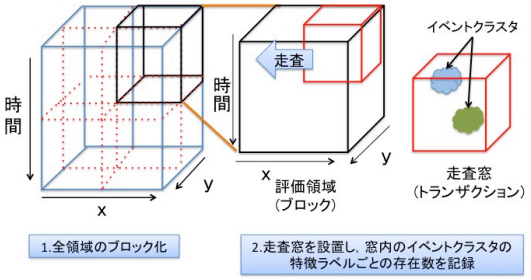


図 2. 時空間データにおけるブロックとトランザクションの定義

(2) 再起性・共起性の評価

走査窓によってサンプリングされたトランザクションをもとに相関ルールにおける支持度と確信度を求めることで再起性・共起性を評価する．相関ルール分析においてデータベースはトランザクションの集合$T=\{t_1,t_2,\dots t_m\}$として表される．ここで各トランザクションは，全アイテム集合 $I=\{i_1,i_2\dots i_k\}$の部分集合によって構成されている．相関ルール$X\rightarrow Y$はトランザクションにアイテム集合Xとアイテム集合Yがともに含まれていることを示す．

相関ルールの支持度は，アイテム集合XとYを含むトランザクションがトランザクション集合全体の中に占める比率で，以下の式で定義される．

$$Supp(X\rightarrow Y)=\frac{|T(X\cap Y)|}{|T|}$$

ここで，$|T(X\cap Y)|$はアイテム集合XとYを含むトランザクションの数を示す．支持度はトランザクション全体にルールを満たすトランザクションがどの程度含まれているかを示している．

確信度は以下の式で表される．

$$Conf(X\rightarrow Y)=\frac{|T(X\cap Y)|}{|T(X)|}$$

確信度は条件を満たしたトランザクションにルールを満たすトランザクションがどの程度含まれるかを示している．時空間データのにおけるイベントクラスタの再起性やイベントクラスタが持つ特徴間の共起性を評価する場合，共起については上記の定義をそのまま利用出来るが，再起については調整が必要である．これについては先行研究[Honda 2015]同様の手法で条件部と結論部を下記の通りに定義して行う．

- 条件部A:ラベルAのイベントクラスタが1つ発生
- 結論部A': ラベルAのイベントクラスタが2つ以上発生

ラベルAのイベントクラスタの再起$A\rightarrow A'$の支持度と確信度は以下の通りになる．

$$Supp(A\rightarrow A')=\frac{|T(A')|}{|T|}\quad Conf(A\rightarrow A')=\frac{|T(A')|}{|T(A)|}$$

共起性を評価する特徴ラベルの組み合わせはアプリオリアルゴリズム[Agrawal 1994]によって得られた多頻度アイテム集合から取り出す．多頻度アイテム集合から取り出した特徴ラベルの組み合わせに対して支持度と確信度を計算し，閾値以上の組み合わせを共起ルールとして抽出する．

再起に関しては領域ごとに閾値以上の支持度・確信度を持つ再起ルールを可視化する．共起性の可視化は抽出されたルールの数をブロックごとにカラーマップで示すことで行う．

3. 実験

相関ルールの指標によるイベントクラスタの再起性・共起性評価手法の実データへの適用例として，時系列気象衛星画像における時間変動の再起性・共起性の評価と可視化を行う．実験には気象衛星 MTSAT-1,2(ひまわり 6,7)が撮影した画像を使用する．図 3 左に画像の諸元を示す．使用した画像は 2012 年 9 月を示す 720 枚の画像で，北緯 70° から南緯 70° ，東経 70° から西経 150° の範囲をマッピングしている．使用画像の一部を図 3 右に示す．この画像の 5×5pixel の輝度の平均値を特徴量とし，112×112×720 グリッドのデータとして評価を行う．評価は図 3 右の赤い線で示される 16 のブロックごとに行うものとする．なお今回は簡単のため時間方向にはブロックを分割せず，31 日を 1 ブロックとして扱った．各トランザクションの収集領域となる走査窓の大きさは X 方向と Y 方向に 7 グリッド，時間方向に 72 グリッドとした．すなわちここでルール抽出の際に近

傍と見なす範囲は緯度経度方向に 140° /112×7=8.75° (赤道で 977km), 時間は 3 日(72 時間)と定義したことになる .

項目	値
画像の種類	MTSAT-1,2IR1
画像サイズ	560×560pixel
画像の範囲	北緯70°～南緯70° 東経70°～東経210°
画像枚数	720枚(2012.8)
空間解像度	0.25°
サンプリング間隔	1時間

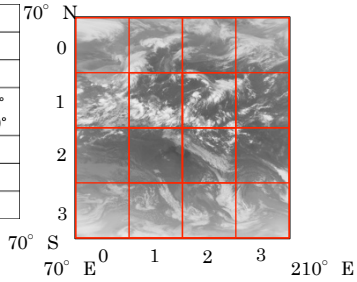


図 3. 使用した時系列画像の諸元と 2012 年 9 月 1 日の画像例とブロックの配置

3.1 部分時系列のクラスタリング

自己組織化マップによる部分時系列学習の際の入力時系列の長さはイベント抽出とイベント長の頻度分布をもとに決定した[森田 2015]. ここではスムージング後の全時系列を閾値によって二値化し, さらに孤立点除去, クロージング処理によって得られたマスク時系列を用いて, イベントを抽出している. 図 4 に対象とする期間に対して求められたイベント長の頻度分布を示す. 自然現象で多く見られる右下がりの指数分布が確認されるとともに, ところどころに頻度が跳ね上がる箇所が見られる. 特に長さ 136 時間(約 5 日間)のイベントは指数分布から予想される値の約 20 倍の値となっており, 特徴的な現象を表す時系列が含まれている可能性がある. そこで, 長さ ω を 136 時間として部分時系列を抽出し, 平均値が 0 になるようオフセットを調整したものを入力としてバッチ型自己組織化マップによる学習を行った.

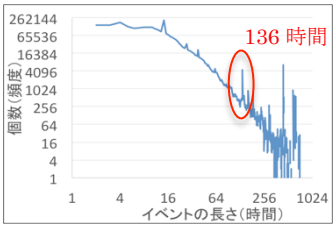


図 4. 2009 年 8 月のイベント長(時間)頻度分布

自己組織化マップの競合層の大きさは 3×3, 近傍距離の初期値は 1, 学習回数は 5 回とした. 図 5 に学習後の競合層の参照ベクトルを表 2 にそれぞれの参照ベクトルの出現頻度と標準偏差を示す. 凡例の番号は競合層のユニットを意味するがこれを部分時系列のラベルと見なす. 4 番の参照ベクトルを見ると変動が小さく出現頻度が一番多いことが確認できる. これは時空間データ中の何も起こっていない非イベントに対応する部分時系列を学習したユニットだと考えられるため 4 番のユニット"以外"に割当てられた部分時系列をイベントとしてイベントクラスタの抽出を行い, 再起性・共起性の評価を行う.

3.2 再起性・共起性の評価と可視化

抽出されたイベントクラスタから時空間のトランザクションをブロックごとに定義した. 再起性, 共起性のそれぞれの評価と可視化は下記のように行った.

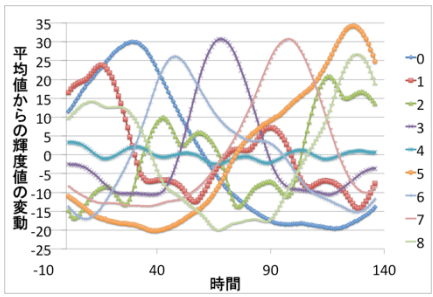


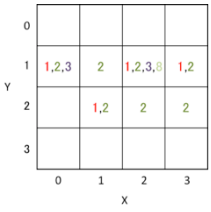
図 5. 自己組織化マップで学習した競合層の重心ベクトル. 部分時系列につけられたラベルの代表例を示す.

表 2. 各ユニットの標準偏差と出現頻度

ユニットID	標準偏差	頻度	出現割合(%)
0	18.47	112119	8.9
1	10.99	128710	10.2
2	11.21	103367	8.2
3	13.25	113474	9
4	1.35	389642	30.8
5	18.47	89481	7.1
6	13.28	126431	10
7	14.91	118543	9.4
8	14.89	85076	6.7

- 再起性の評価

支持度の閾値を 0.3, 確信度の閾値を 0.6 として再起ルールの抽出を行った. 図 6 に各ブロックの閾値以上の支持度, 確信度を持つルールを可視化したものを示す. 各ブロックの数字は再起ルールが抽出されたユニットの Id である. 図 6 を見ると ラベル 1,2 のイベントクラスタの再起が赤道付近のブロック(2,1)で多数発生していることが分かる.



Rule	Supp	Conf
[1]->[1]	0.44	0.635
[2]->[2]	0.612	0.778
[3]->[3]	0.443	0.615
[8]->[8]	0.414	0.649

図 6. 抽出された再起ルールごとの空間分布(左)と赤丸でしめしたブロック(2,1)で抽出されたルールの詳細(右)

図 7 にユニット 1, 2 の参照ベクトルを示す. これは 5 日間のうちに2回のピークあるような時系列であることが分かる. この結果から, ルールが抽出されたブロックで 5 日間の間に雲が断続的に発達しては消滅するような現象が発生した場合 3 日以内に近傍で同様のイベントが支持度 0.3, 確信度 0.6 で発生することを意味している. なお, 1 と 2 は同じ現象の位相ずれに見えるため, これについては FFT パワースペクトル等位相の影響のない特徴を使用することが必要であろう.

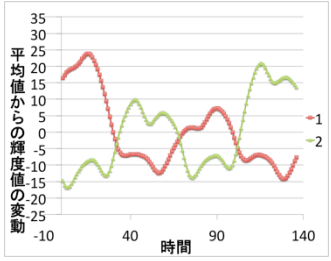


図 7. ユニット 1, ユニット 2 の参照ベクトル

- 共起性の評価

アプリオリアルゴリズムによって得られた多頻度アイテム集合から支持度 0.3, 確信度 0.8 以上のユニット集合の組み合わせを

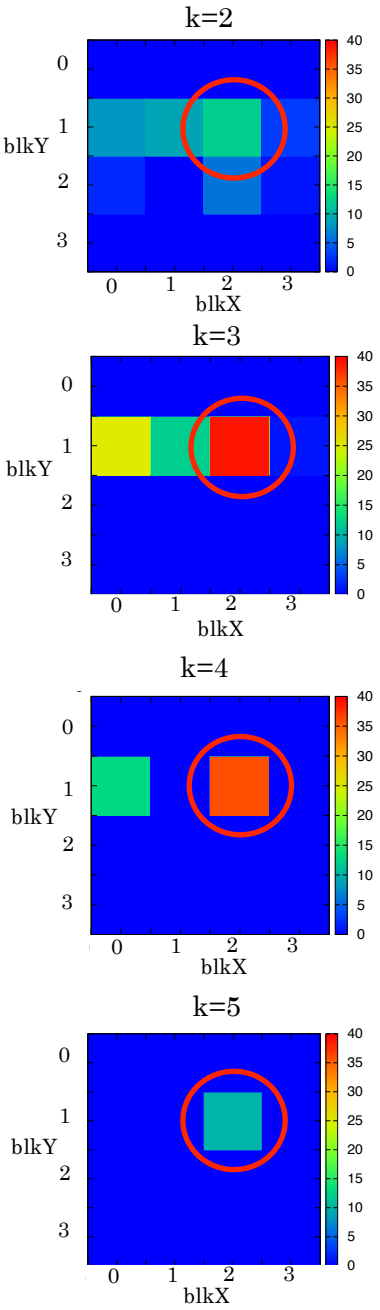
共起ルールとして抽出した. 図 8 に抽出された多頻度アイテム集合の項数 k ごとの抽出ルール数をカラーマップで示す. 抽出されたルールは 2-5 項からなるものであったが, どの項数の共起ルールも再起ルールと同様に赤道付近のブロックで多く抽出されていることが分かる. 特に上から 2 番目左から 3 番目のブロック(2,1)ではどの項数のルールも多数抽出されていることが分かる. 図 8 の右側の表はブロック(2,1)で抽出された共起ルールの一覧とその支持度と確信度である. 抽出されたルールを見ると, 結論部にユニット 2 のイベントクラスタが多く現れていることが分かる. 図 6 の再起の分布図からもこのブロックでユニット 2 に割当てられたイベントクラスタの再起性の高さが確認できることから, このブロック内のトランザクションのほとんどに[2]のイベントクラスタが含まれており, 多くのイベントクラスタとの共起が抽出されているということが考えられる.

例えばブロック(2,1)で抽出された項数 k=2 のルール[5]→[2](パターン図 5 参照)は, このルールの支持度は 0.44, 確信度は 0.921 であった. これは, このブロック内の全トランザクションの約 4 割でこのルールが確認できること, ユニット 5 の特徴を持つイベントクラスタが発生した場合に 5° ×5° の近傍で前後 3 日以内に 92%の割合でラベル 2 の特徴を持つイベントクラスタが発生することを意味する. このように, どのような領域でイベントクラスタの共起が起こりやすいかを視覚的にとらえたうえで, 支持度と確信度を用いて共起ルールを定量的に評価することが可能となった.

4. おわりに

時空間データに存在する多様な時系列変動から自律的に特徴を要約し, それぞれの特徴を持つイベントの塊であるイベントクラスタの再起性と共起性を時空間の領域ごとに定量的に評価することで, その時空間分布を可視化する手法を検討した. 先行研究では注目すべき基準時系列を与え, 基準と相関の高いイベントクラスタ同士の再起のみをあつかっていたが, 自己組織化マップを用いて特徴を自律的に要約することにより, 異なる特徴間の共起も扱うことが可能となった. 各特徴間の共起性評価の際は, アプリオリアルゴリズムを時空間のトランザクションに対して適用し効率的に評価を行った. 再起性・共起性は相関ルールにおける支持度と確信度を用いて定量的に評価し, 領域ごとのルールを視覚化することで, 再起・共起パターンの空間分布を提示した.

なお, 今回の手法では時空間のトランザクションを走査窓として定義しスライドさせながらイベントクラスタの発生順序に関係なく共起を抽出し評価した. 時空間データにおいては時間に依存したパターンが重要であると考えられるため, 時間依存のパターンの評価を行うためのトランザクションの定義方法を今後検討していく予定である.



Rule	Supp	Conf
[5]->[2]	0.44	0.921
[7]->[2]	0.536	0.919
[8]->[2]	0.573	0.898
[1]->[2]	0.622	0.896
[3]->[2]	0.64	0.888
[6]->[2]	0.562	0.886
[7]->[3]	0.504	0.865
[8]->[3]	0.544	0.852
[6]->[3]	0.535	0.844
[5]->[3]	0.389	0.815

Rule	Supp	Conf
[5,7]->[2]	0.321	0.94
[1,7]->[2]	0.433	0.931
[1,5]->[2]	0.339	0.93
[5,6]->[2]	0.337	0.928
[6,7]->[2]	0.408	0.924
[3,5]->[2]	0.359	0.922
[3,7]->[2]	0.463	0.919
[7,8]->[2]	0.416	0.918
[5,8]->[2]	0.324	0.918
[1,8]->[2]	0.465	0.918

Rule	Supp	Conf
[1,7,8]->[2]	0.343	0.934
[1,6,7]->[2]	0.335	0.931
[1,3,7]->[2]	0.375	0.929
[6,7,8]->[3]	0.325	0.928
[1,3,8]->[2]	0.403	0.922
[6,7,8]->[2]	0.323	0.92
[3,7,8]->[2]	0.375	0.92
[3,6,7]->[2]	0.363	0.919
[1,6,8]->[2]	0.345	0.919
[3,6,8]->[2]	0.374	0.909

Rule	Supp	Conf
[1,3,7,8]->[2]	0.31	0.934
[1,3,6,8]->[2]	0.314	0.921
[1,2,6,8]->[3]	0.314	0.91
[1,2,7,8]->[3]	0.31	0.904
[1,7,8]->[2,3]	0.31	0.845
[2,3,6,8]->[1]	0.314	0.838
[1,6,8]->[2,3]	0.314	0.837
[2,3,7,8]->[1]	0.31	0.826
[1,2,3,7]->[8]	0.31	0.825
[1,2,3,6]->[8]	0.314	0.81

図 8. 各ブロックにおいて抽出された項数 k の共起ルールの個数分布(左). カラーバーは抽出されたルールの個数を示す. 右の表は, 例として(2,1)(赤丸)のブロックについて抽出された確信度上位 10 個のルールを示す.

参考文献

[Matsubara 2014] Matsubara S., Sakurai, S. , G van Pandhuis, W. and Faloutsos, C.: FUNNEL: Automatic Mining of Spatially Coevolving Epidemics , ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, p. 105-114(2014).

[Kohonen 1990] Kohonen, T.: Self-Organizing Maps 3rd ed., Springer Verlag, pp. 528(2000).

[何 2009] 何立風, 巣宇燕, 鈴木賢治, 中村剛士, 伊藤英則: 三次元 2 値画像における高速ラベル付けアルゴリズム, 電子情報通信学会論文誌 D, 92.12, p.2261-2269(2009).

[Agrawal 1994] Agrawal, R., and Srikant, R.: Fast algorithms for mining association rules, Proc. 20th int. conf. very large data bases, VLDB. Vol. 1215, p. 478-499(1994).

[森田 2015] 森田博次:分散処理フレームワークを用いた時系列データからのイベント検索システムの構築, 高知大学理学部応用理学科卒業論文(2015)

[Honda 2015] Honda, R., and Mori, K.: Extraction of Highly Correlated Temporal Event Cluster Recurrence from Spatiotemporal Data, In 2015 IEEE International Conference on Data Mining Workshop (ICDMW), p. 1457-1461(2015)