Temporal Skeltonization on Sequential Data – Patterns, Categorization, and Visualization –

C. Liu, K. Zhang, H. Xiong, G. Jiang, Q. Yang

KDD2014読み会

大滝啓介 (京都大学)
@cocomoff

(all figures are from the paper)

概要

Pattern Miningの実応用における諸問題を扱う

- 列パターンマイニング
  - ログ解析や軌跡解析を系列パターンマイニングとしてモデル化することは，そこまで簡単ではない
  - 記号列化するのが難しすぎる

手法

- Temporal Skeltonization ~ 記号のクラスタリング
- Spectral clusteringの応用

実験と結果

- 人工データと実データ(B2B購買履歴データ)
- いい感じに抽象的な解析ができた

Sequential Pattern Mining

$S = \{e_1, \ldots, e_n\}$ 上の系列を $\mathrm{S_i} = \left(s_1^i, \ldots, s_{T_i}^i\right) \in S^{T_n}$ と書く。データベースを $DB = \{\mathrm{S}_1, \ldots, \mathrm{S}_N\}$ と定義し、列の評価関数をqと書く。パラメータθに対して $q(DB, \mathrm{S}) \geqq \theta$ であるような列Sをすべて列挙せよ。

- 定義に関してはVariationがある
  - 例)Apriorで有名なAgrawal & Srikantコンビの論文(1995)における定義はこれではなかったはず。

関連研究・ソフトウェア

- **論文:** GSP(Agrawal&Srikant,96), SPADE(Zaki,01), PrefixSpan(Han et al.,01), SPAM(Ayres et al., 02)
- **使いたい場合**: RのarulesSequences, PrefixSpanの実装(←たくさんある), SPMF(Javaベース)など

何が問題か

- データ(購買履歴・軌跡・イベントログ)を系列データベースに変換するには大きなΣが必要
  - 英単語文章 $|S| \fallingdotseq 10{,}000$, 購買履歴 $|S| \rightarrow$ すごく大きい
- Σが大きくなると**アルゴリズム的には嬉しくない**
  - 探索空間が大きくなる(最悪探索不可能に)
  - 評価値がどれも似たような感じに(スパース化)
  - 価値のある情報か分からない(ノイズ)

その他の事情

- 記号の階層的背景知識って有効だろう
  - 階層的な知識ってどうやって用意するんだ
  - 記号から意味情報が落とされている場合はどうする
  - プライバシー保護データマイニング的の問題点
- 記号の関係を自動的に考慮できると嬉しい

例

Graph的知見に基づいたイベント記号のクラスタ化

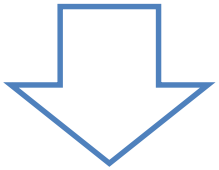
- Skelton ~ Temporal cluster
  - Temporal pattern: series of events, time-series data
  - 系列的に局所的かつ系列的なクラスタのこと？

例

- Σ={a,b,c,…,o}上の系列に対し，Σを3分割する

$$1 : \mathtt{m,h,j,f,d,a,i,k,b}$$
$$2 : \mathtt{j,l,m,a,n,f,b,o,g}$$
$$3 : \mathtt{e,h,l,c,f,n,i,b,o}$$
$$4 : \mathtt{h,l,e,c,a,f,k,o,i}$$

$$\mathtt{A} = \{\mathtt{m,h,j,e,l}\}$$
$$\mathtt{B} = \{\mathtt{a,f,c,d,n}\}$$
$$\mathtt{C} = \{\mathtt{k,g,o,b,i}\}$$



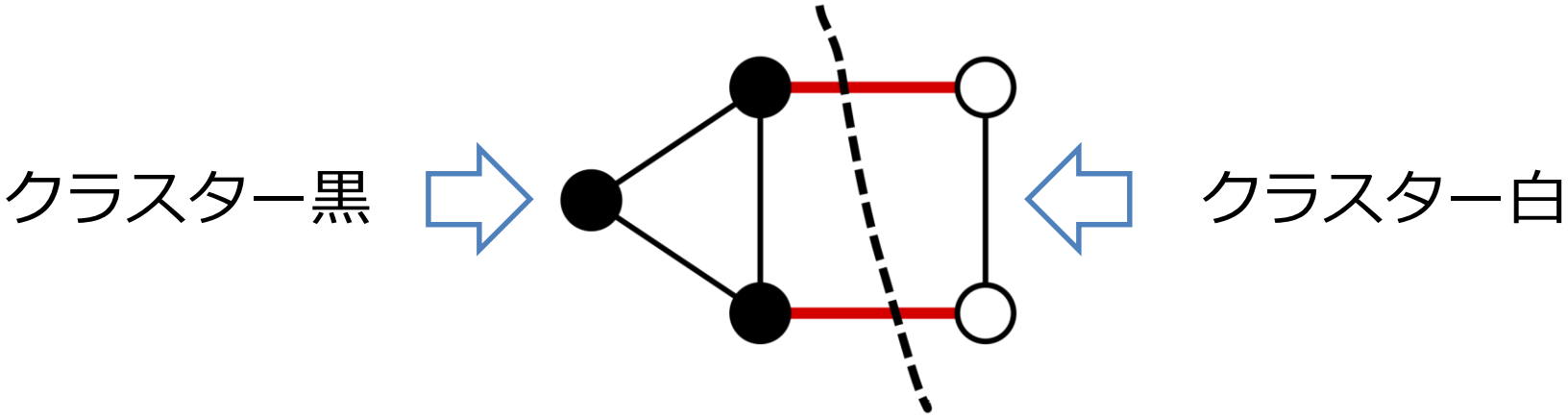
関数 f: S→{1,2,…,K} (Encoding scheme)

A, A, A, B, B, B, C, C, C

例) |S|=20 → K=3

Graph Cut & Spectral Clustering

- グラフ切ります (Wikipediaより図)

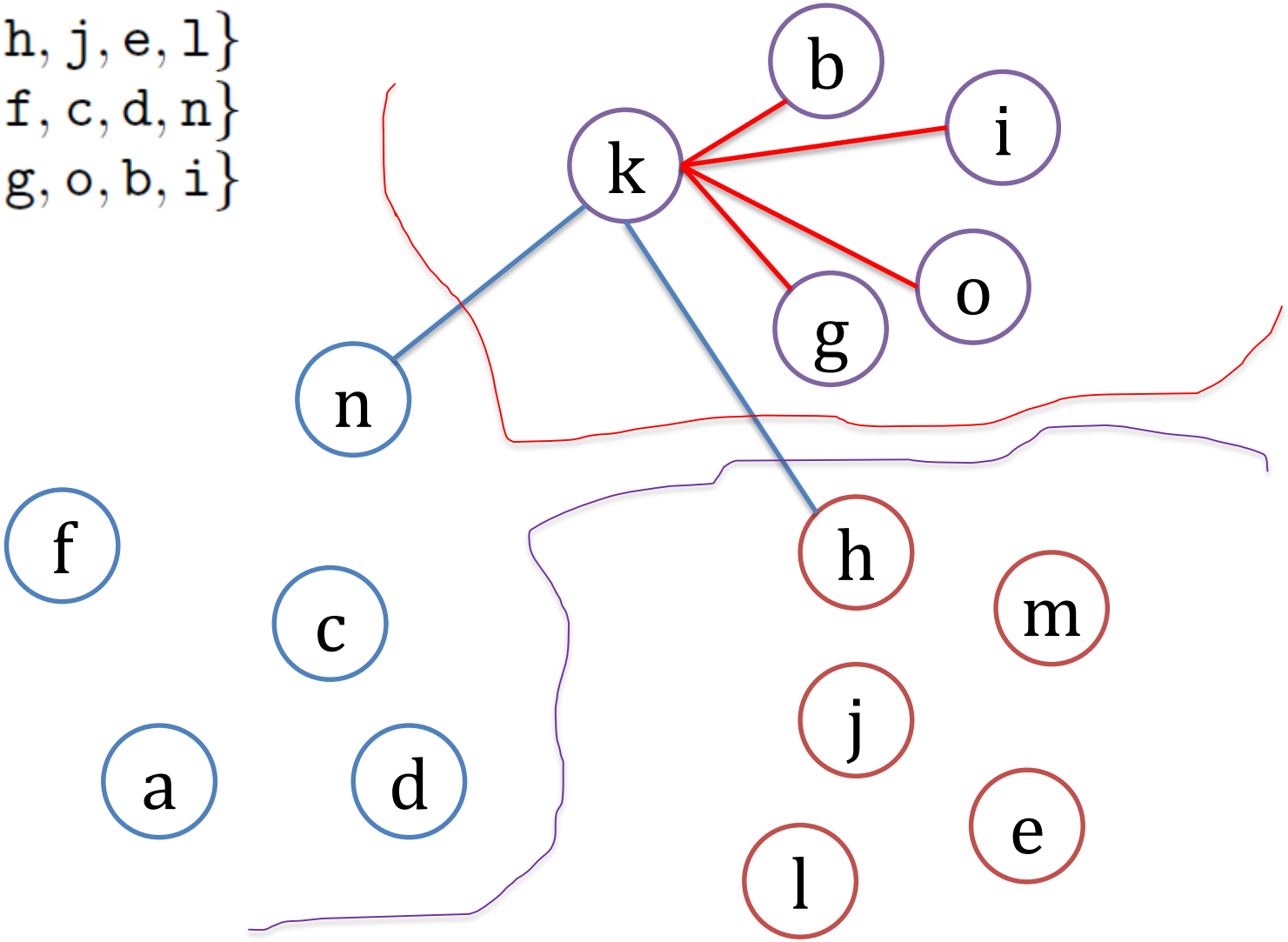


- Spectral clustering
  - 接点にデータ，辺にpair-wiseな類似度を与える
  - ラプラシアン行列 L = D – Wを計算する
    - Dは次数行列，Wは重み行列
  - Lを固有値分解して固有値の大きい方を取ってくる
  - Graph-cutと関係があり
  - 詳しくはgoogleで「Spectral clustering」を検索

例

- Encoding scheme f を取得する問題とは

$$A = \{m, h, j, e, l\}$$
$$B = \{a, f, c, d, n\}$$
$$C = \{k, g, o, b, i\}$$



Temporal Graph

DEFINITION 1 (TEMPORAL GRAPH). *Let G be a weighted graph $G = \langle V, E \rangle$ with vertex set $V = \mathcal{S}$ and edges E. The i-th node of G corresponds to the i-th symbol e_i in the symbol set $\mathcal{S}$. The weight of the edge between node i and node j is defined as the ij-th entry of an $|\mathcal{S}| \times |\mathcal{S}|$ matrix W, where*

$$W_{ij} = \frac{1}{N} \sum_{\substack{1 \le n \le N \\ e_i, e_j \in S_n}} \delta\left(|\ell(e_i, S_n) - \ell(e_j, S_n)| \le r\right). \qquad (2)$$

Here, we say $e \in S_n$ if the symbol e can be observed in the sequence S_n and $\ell(e, S_n) \in \{1, 2, \cdots, T_n\}$ is the corresponding location of e in the sequence S_n.

イメージ)データベース中で近くに出現する
アルファベットを大きくカウントする

Temporal Skeltonization Problem

PROBLEM 1 (TEMPORAL SKELETONIZATION). *Given a set of sequences* $\{S_n | n = 1, 2, \cdots, N\}$, *we want to find a new encoding scheme of the symbols* $e \in \mathcal{S}$, *denoted by the mapping* $y = f(e) \in \{1, 2, \cdots, K\}$, *such that when encoded with* f, *the temporal variation of resultant sequences is minimized*

$$\min_{y \in \{1,2,\ldots,K\}} \frac{1}{N} \sum_{n=1}^{N} \sum_{\substack{1 \le p,q \le T_n \\ |p-q| \le r}} \left(f(s_p^n) - f(s_q^n) \right)^2. \qquad (1)$$

Here r *is a pre-defined integer that controls the range that local sequence variations are computed, and the cardinality of the encoding scheme,* K, *is a pre-defined integer that is much smaller than that of the original representation* $|\mathcal{S}|$.

Temporal Skeltonization Problem

- グラフ表現により以下のように定式化される
  - 式3: $y_i = f(e_i)$, coding scheme f and character e

$$\min_{\mathbf{y}\in\mathbb{R}^{|\mathcal{S}|}} \sum_{i,j} W_{ij}(\mathbf{y}_i - \mathbf{y}_j)^2.$$

- 以下と同じ形(式4)
  - これはspectral clustering, graph-cutsで頻出するグラフベースの最適化問題によく出てくる形
  - 解ける

$$\min_{\mathbf{y}\in\mathbb{R}^{|\mathcal{S}|}} \quad \mathbf{y}'(D-W)\mathbf{y}$$

$$s.t. \quad \mathbf{1}'D\mathbf{y} = 0$$

$$\mathbf{y}'D\mathbf{y} = 1$$

実験 (1) 人工データ

- 5 groups {A,B,C,D,E}, each of them has 25 symbols
  - 5000 sequences
    - 2500: A→B→C→D
    - 2500: B→E→C
      - 各グループでランダムにd個記号を選択
      - 平均で長さが15になるように調整

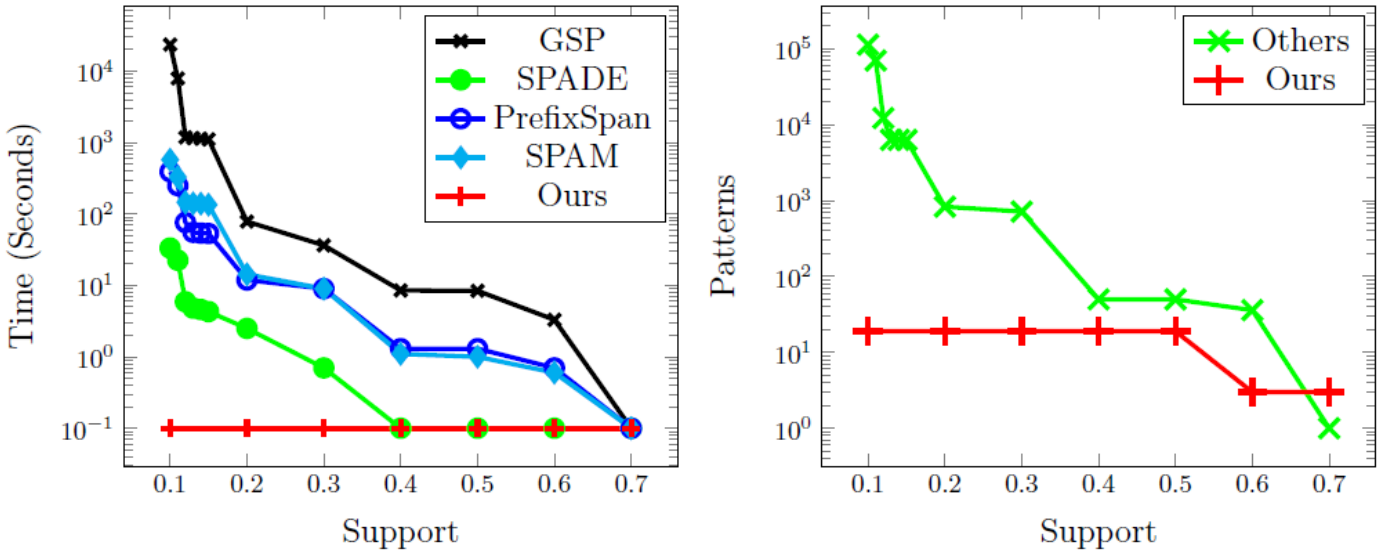
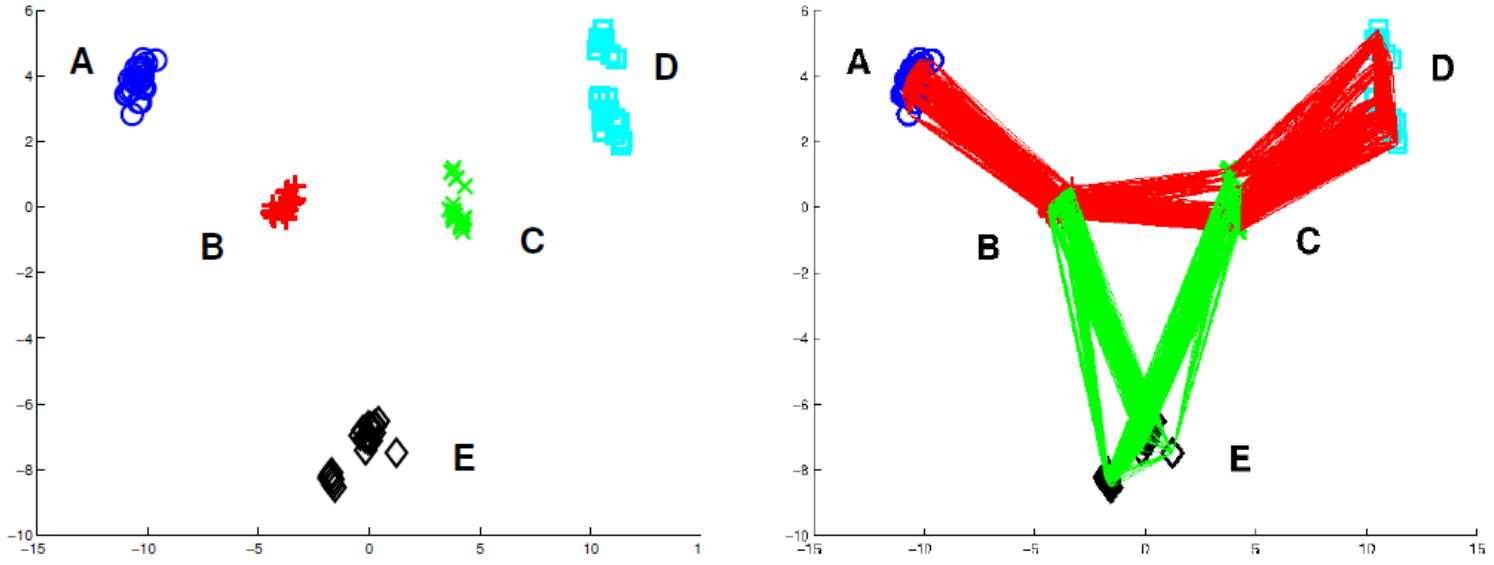


Figure 3: FSM algorithms on the simulated data.

実験 (1) 人工データ

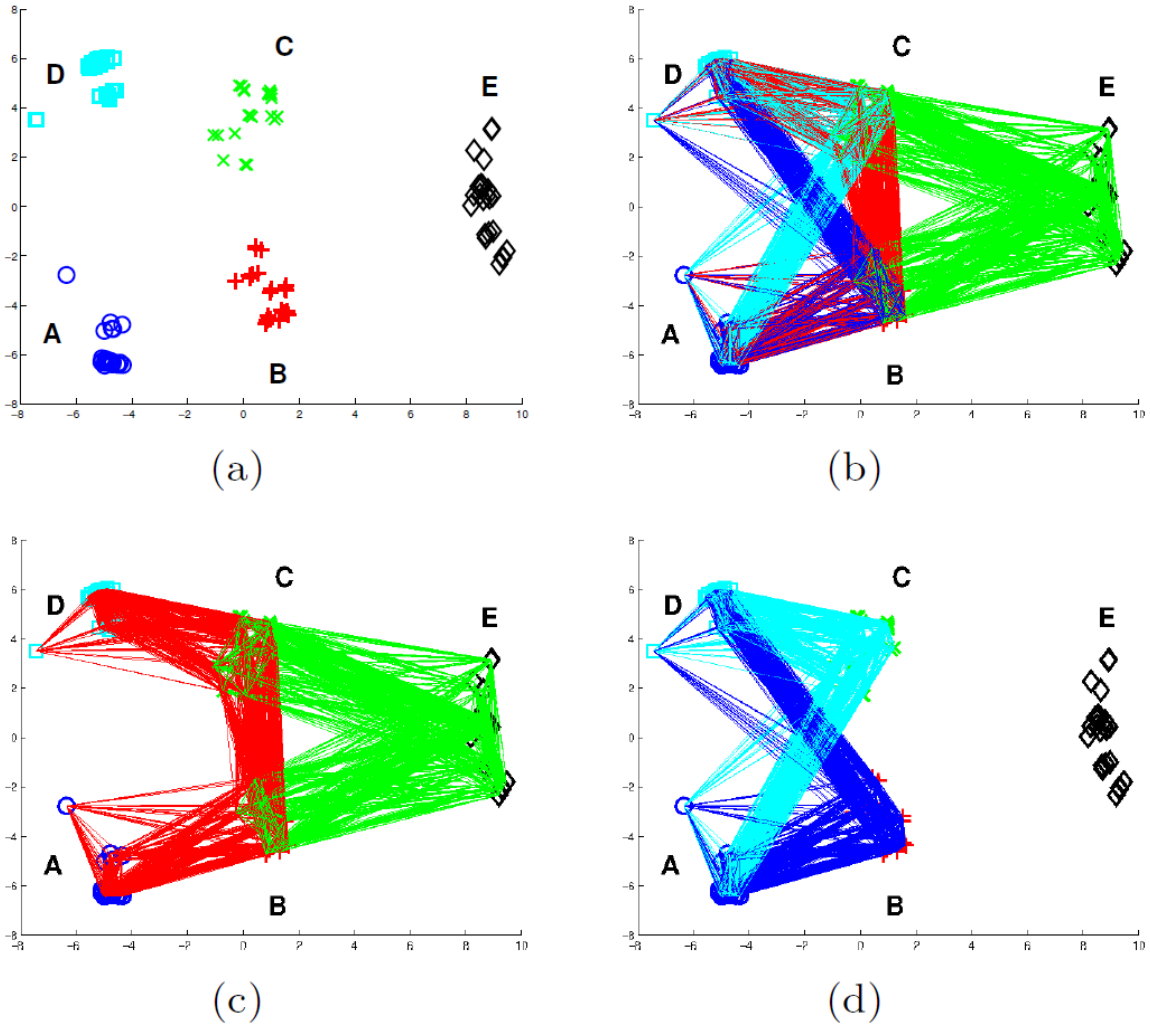
- 5 groups {A,B,C,D,E}, each of them has 25 symbols
  - 5000 sequences
    - 各グループでランダムにd個記号を選択
    - 平均で長さが15になるように調整
  - 可視化のためにIsomapを使っているらしい



(a) The temporal clusters. (b) The sequence clusters.

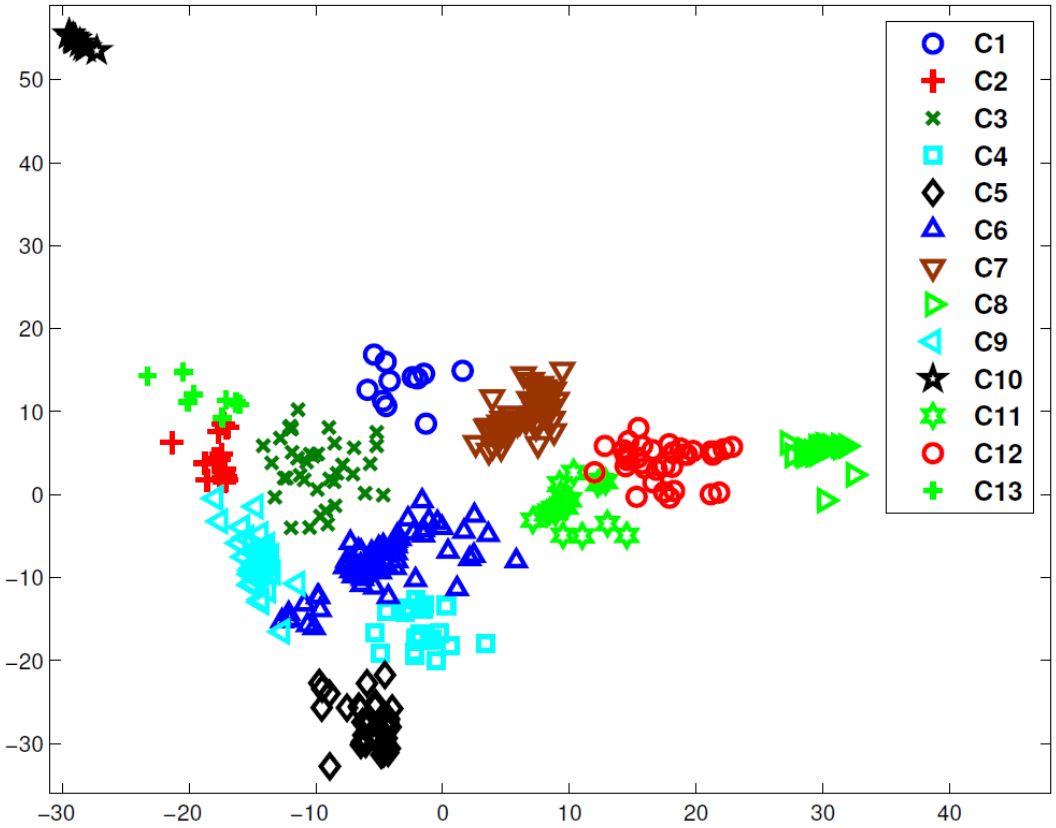
実験 (1) 人工データ

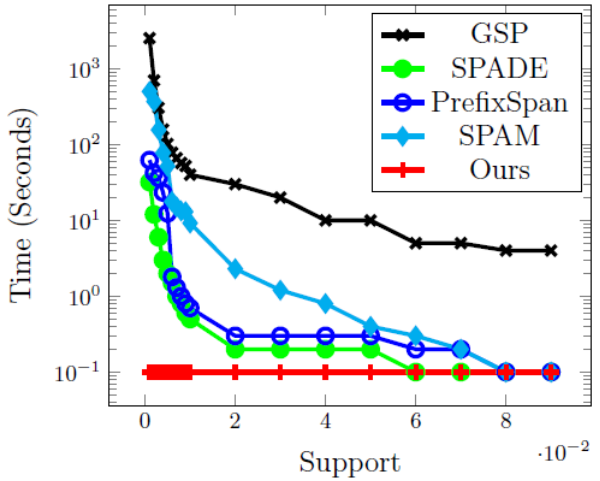
- ノイズあり(別のgroupの記号を5%入れる)

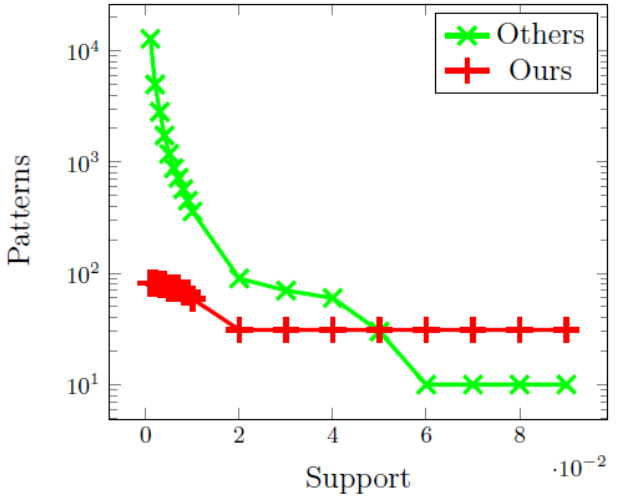


実験(2) 実データ

- $|S|=5028, T=\sum_n T_n = 248725$







実験(2) 実データ

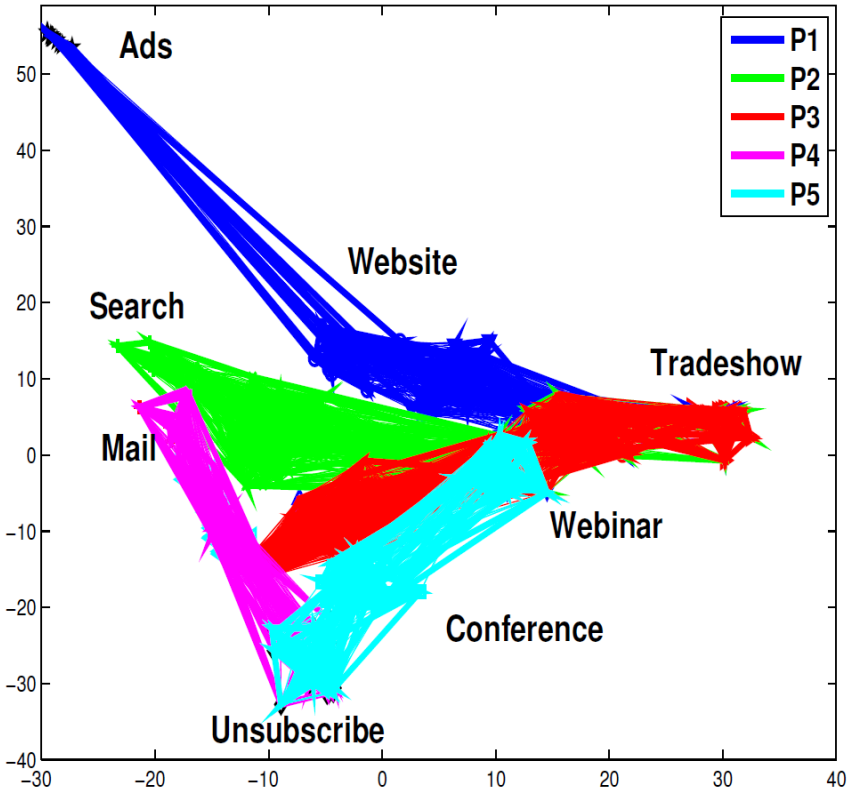
- 実際にクラスタを見てみるとそれなりに上手くアノテーションされた形になるらしい

Cluster	Top keywords	Size
C_1	Official Website	12
C_2	Corporate Event, Direct Marketing Mail	20
C_3	Trial Product Download	45
C_4	Conference	27
C_5	Unsubscribe	38
C_6	Webinar	101
C_7	Trial Product Download	70
C_8	Tradeshow	37
C_9	Corporate Event, Direct Marketing Mail	65
C_{10}	Web Marketing Ads	13
C_{11}	Webinar	21
C_{12}	Webinar	42
C_{13}	Search Engine	12

Table 3: The semantic annotation of event clusters.

実験(2) 実データ

- 通常の記号より抽象的なクラスタラベルを使った方が様々な解析がしやすいのではないか



Class	Path	Path/Keyword	Size
Successful	P_1	$C_{10} \rightarrow C_1 \rightarrow C_7 \rightarrow C_{12} \rightarrow C_8$ Ads→Website→Download→ Webinar→Tradeshow	933
	P_2	$C_{13} \rightarrow C_3 \rightarrow C_{11} \rightarrow C_{12} \rightarrow C_8$ Search→Download→Webinar→ Webinar→Tradeshow	1110
	P_3	$C_6 \rightarrow C_{11} \rightarrow C_{12} \rightarrow C_8$ Webinar→Webinar→Webinar→ Tradeshow	702
Unsuccessful	P_4	$C_2 \rightarrow C_9 \rightarrow C_5$ Mail→Corporate Event→ Unsubscribe	423
	P_5	$C_{11} \rightarrow C_4 \rightarrow C_5$ Webinar→Conference→ Unsubscribe	333

まとめ

より可読性の高いパターンマイニングを目指して

- 記号集合をクラスタ分割する前処理を行う
  - 記号の書き換えスキーマ Encoding scheme f
  - Temporal Graphを構築，Spectral clusteringを行う
  - 発見されたクラスタを用いた系列データベースのスムーズ化・エンコーディングについて(飛ばした)
- いくつかの実験結果を示した
  - ノイズがあってもクラスタ化が可能(人工データ)
  - なんとなくそれなりにわかりやすい(実データ)
- その他の応用について
  - パターンの可視化，パターンマイニングの改良
  - 基本的に系列データ以外にも適用可能