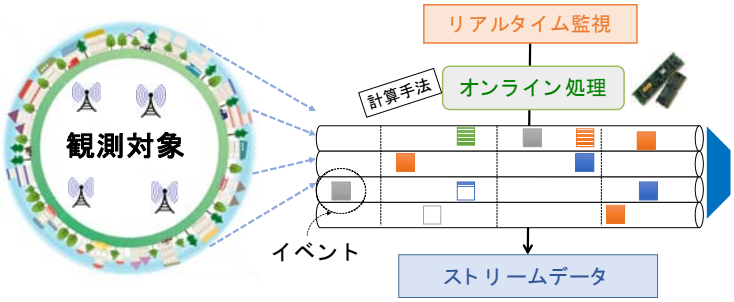
静岡大学

ストリームデータの要約技術とその応用

情報学領域　情報科学系列　准教授　山本泰生

ストリームデータの研究

- ストリームデータとは？
  - 高速に流れ続ける無限長のデータ列
  - センサーノードから常時到着する観測データ
  - 観測対象のリアルタイム分析 (傾向の変化や異常の検出)



ストリームデータのモデリング

- 半順序 (束) 関係に基づく代数モデル
  - 解析対象のイベントの全体集合: E
  - ストリームデータ: $V_n = \langle e_1, e_2, \ldots, e_n \rangle$ ただし $e_i \in E$
  - E 上のある半順序(束)関係 を想定する (以降，関係を R と書く)
- サポートとクエリ
  - イベント e の**サポート**: $\mathrm{R}(e, ei)$ を満たす V_n 中の ei の個数

イベント e と関係を結ぶ (「サポート」する) イベントの発生件数のこと

気象データ (関係データベース)

V_3	気温 a_1	降水量 a_2	風速 a_3	風向 a_4	日照 a_5
e_1	14	5	3	0	10
e_2	11	10	1	0	5
e_3	19	5	2	1	14

各イベント e は 5属性 (気温, 降水量, 風速, 風向, 日照) から構成される; $e_i = (a_1^i, a_2^i, a_3^i, a_4^i, a_5^i)$ と書く

「からから」 関係 R を次のように想定

任意の $e_i, e_j \in E$ に対し

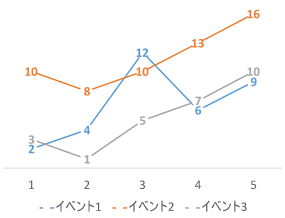
$R(e_i, e_j)$ 定義 $\leftrightarrow$ $a_1^i \le a_1^j$ かつ $a_2^i \ge a_2^j$

e_i より e_j の方が からから　気温が高く　降水量が低い

Q. 気温 12 ℃ 以上 & 降水量 5 mm 以下は何日?

Q. 誤差 3 の範囲でパターン (3, 4, 5, 6, 7) に含まれる時系列は過去に出現した？

時系列データ
- 測定値には必ず誤差が含まれる
- 「区間」として測定値を表現する



	e1		e2		e3	
時刻	MIN	MAX	MIN	MAX	MIN	MAX
1	2	2	10	10	3	3
2	4	4	8	8	1	1
3	12	12	10	10	5	5
4	6	6	13	13	7	7
5	9	9	16	16	10	10

サマリとは？

ストリームデータ: $V_n = \langle e_1, e_2, \ldots, e_n \rangle$ ただし $e_i \in E$ とする

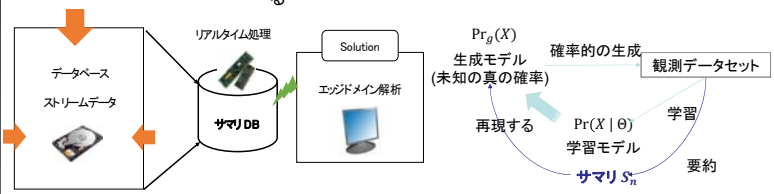
サマリ とは
任意のイベント e に対し，V_n における e のサポートを与えるデータ構造
クエリに相当する

サマリの構築法

- 元のデータベースそのまま
- 可逆圧縮 (きちんと管理)
  - 空間計算量 & 時間計算量は良くて O(n) (n はデータ量)
- 非可逆圧縮 (ゆるく管理)
  - 劣線形計算量 ~ O(log n) ~ を目指す (超軽量 & 超高速！)
  - リアルタイム解析に適したオンラインアルゴリズム

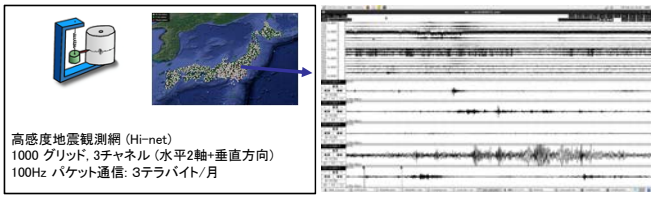
非可逆圧縮サマリのインパクト

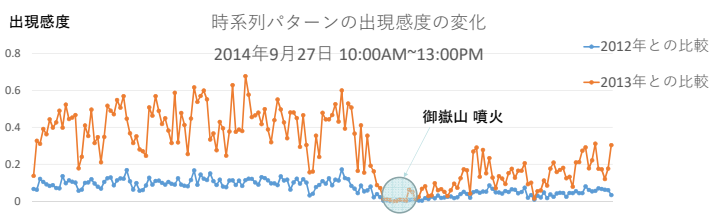
以下の3つの条件を満たすサマリDBを構築
(1) **一度きり**のデータスキャンで良い
(2) **メモリ内に保持できる**
(3) サポートの見積値の**誤差保証**ができる



サマリの用途 (何ができる？)

- 想定する関係を満たすイベントが過去に出現したかどうか，もし出現していれば何回出現しているか教えてくれる
- 近接事例ベース異常検出 (proximity-Instance-based anomaly detection)
  - 事前の学習モデルを必要としない
  - ノンパラメトリック (=解析の対象データに一切の確率分布を仮定しない)
    → 現象の発生メカニズムがわからないデータに利用可能
  - モデルのダイナミクスが時間変化する概念遷移に対応可能





keyword:　ビッグデータ管理・リアルタイム解析・データ駆動発見

■ お問合せ ■　静岡大学　イノベーション社会連携推進機構　(ＴＥＬ)053-478-1702 (代) (E-mail) sangakucd@adb.shizuoka.ac.jp