

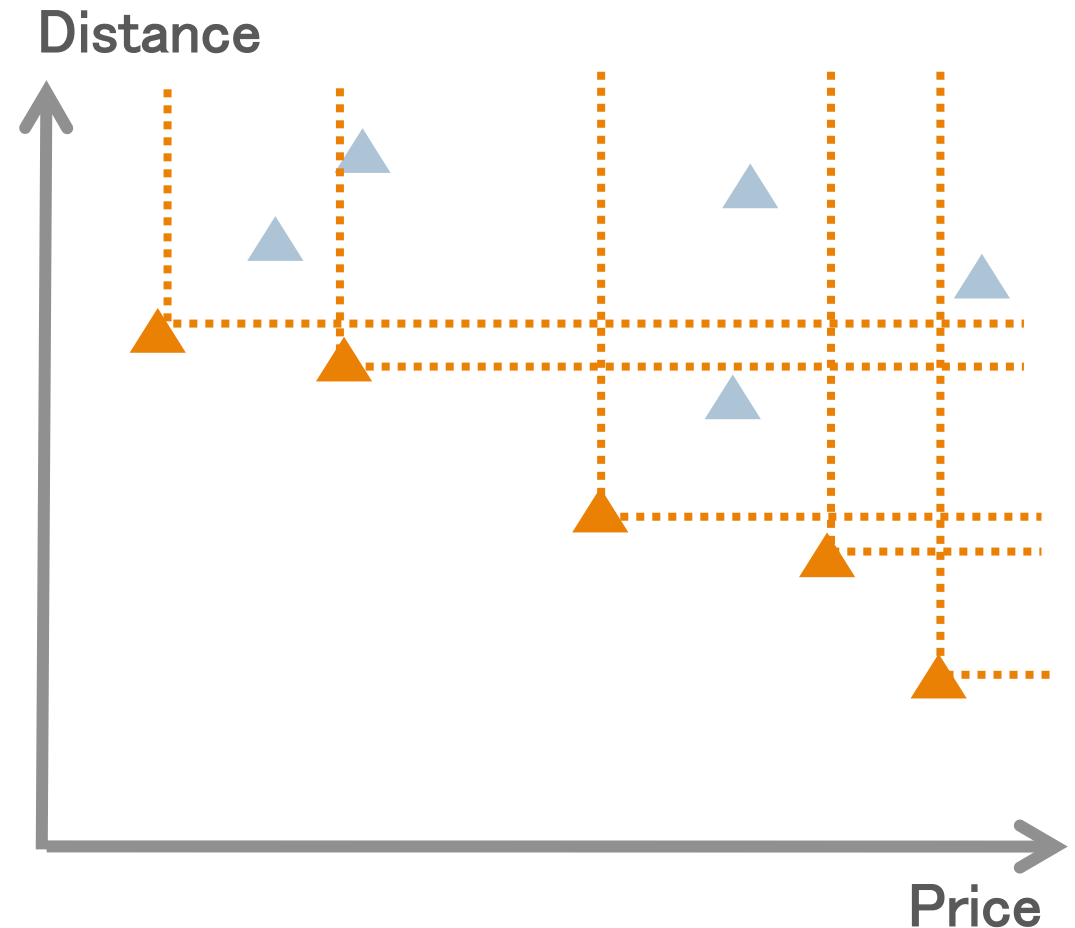
【VLDB2011勉強会】

Session 30: Skyline and String Matching

担当：天笠俊之（筑波大学）

Skyline問合せとは？

- ▶ 支配 (dominance)
 - ▶ 点pが点qに対してすべての次元で優位であるとき、pはqを支配するという。
- ▶ 問題
 - ▶ 与えられた多次元データ点集合から、他のデータ点に支配 (dominate) されない点を列挙する。



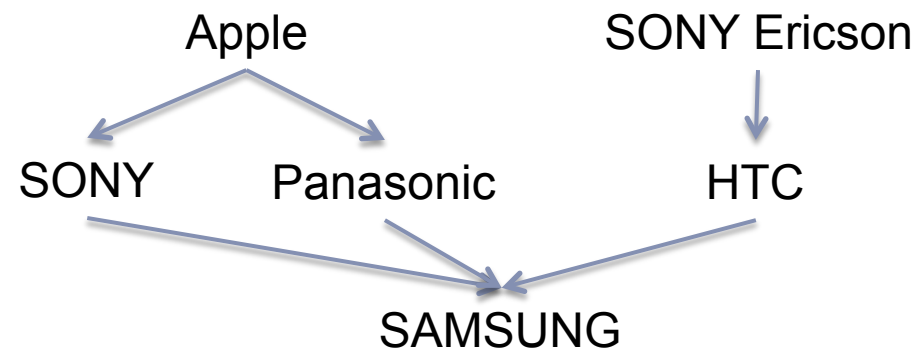
ZINC: Efficient Indexing for Skyline Computation

▶ 著者

- ▶ Bin Liu, Chee-Yong Chan (SNU, Singapore)

▶ 概要

- ▶ ZINC (Z-order Indexing with Nested Code) を提案.
- ▶ ZB木と入れ子エンコーディングの組合せ.
 - ▶ 半順序 (Partial Order; PO) をサポート.
 - (時)区間, 型階層, 集合比較などに応用可能.



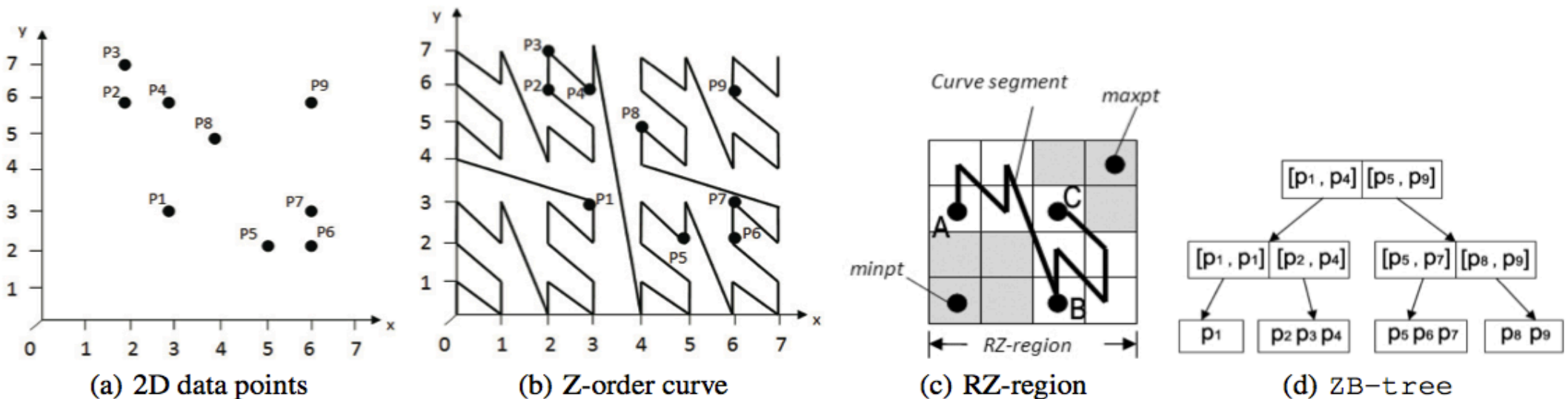
前提知識: ZB木

▶ Zカーブ

- ▶ 空間を四つの領域に分割.
- ▶ Zカーブ上で先に出現するデータ点は後に出現するデータ点に支配されない.

▶ ZB木

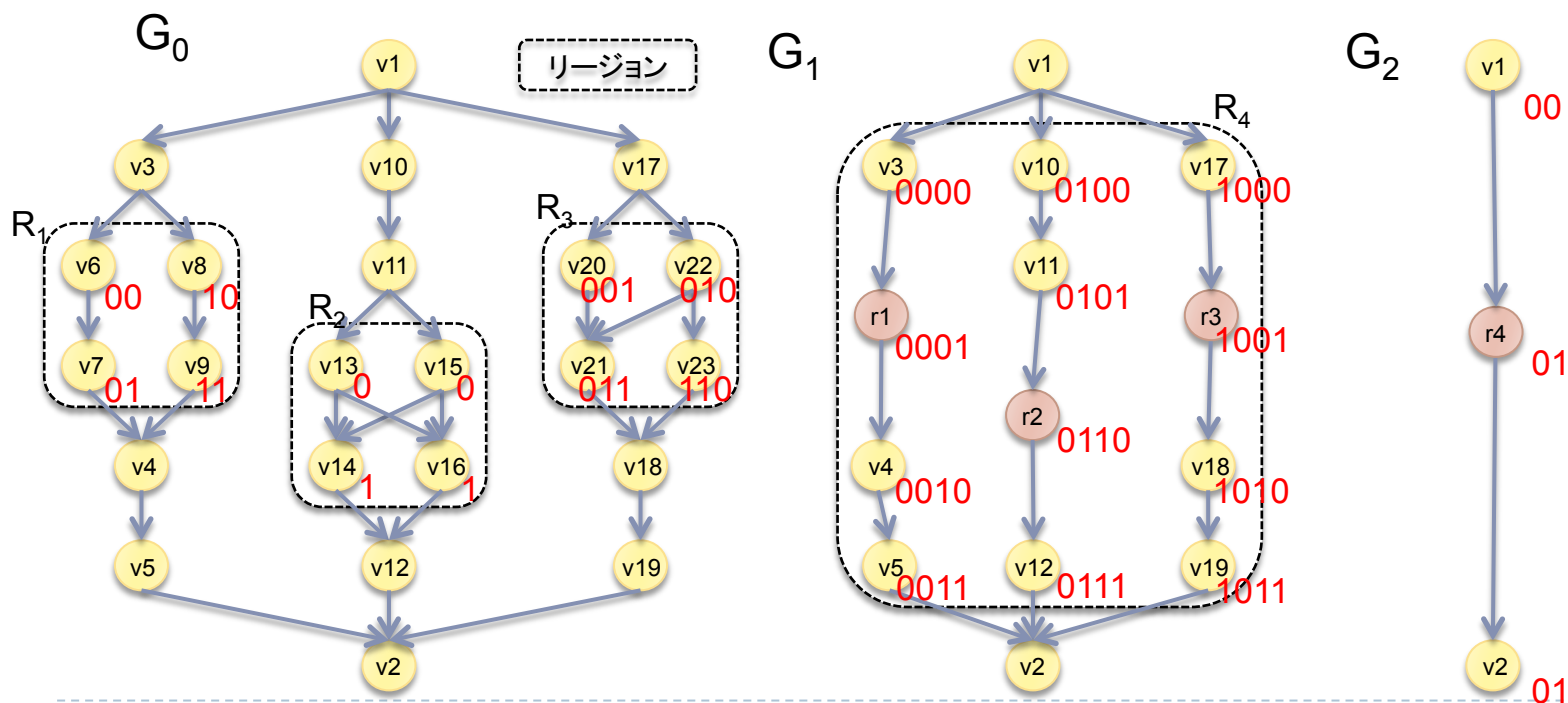
- ▶ 点を, Zアドレスを用いてB+木に格納.



入れ子エンコーディングによる半順序のコード化

▶ アプローチ

- ▶ 順序束において、親(子)を共有する部分グラフをリージョンとして取り出し、ノードに縮約。
 - ▶ Horizontal / Vertical / Irregular region
- ▶ リージョン内の順序関係を再帰的にコード化。



$$\begin{aligned} \text{Encode}(v_1, G_0) &= \text{Encode}(v_1, G_2) \\ &= \underline{00\ 0000\ 000} \end{aligned}$$

$$\begin{aligned} \text{Encode}(v_{20}, G_0) &= \text{Encode}(r_4, G_2) \\ &+ \text{Encode}(r_3, R_4) \\ &+ \text{Encode}(v_{20}, R_3) \\ &= \underline{01\ 1001\ 001} \end{aligned}$$

評価実験

▶ データセット

▶ 合成データ

- ▶ [D.Sacharidis, ICDE'09]

▶ 実データ

- ▶ Netflix, MovieLensを変換

▶ 比較手法

▶ TSS

- ▶ [D.Sacharidis, ICDE'09]

- ▶ PODメインに対応.

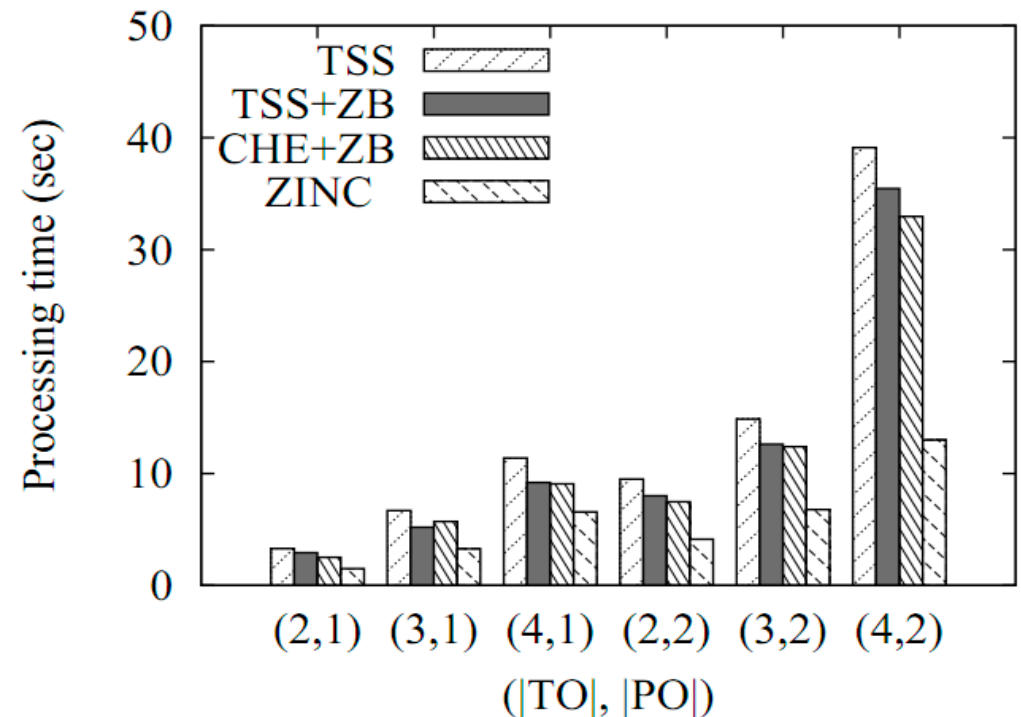
▶ TSS+ZB

- ▶ TSSの符号化法 + ZB木

▶ CHE+ZB

- ▶ 多重継承の符号化法[Y.Caseau, OOPSLA'93] + ZB木

▶ ZINC



著者の発表資料より引用:

<http://www.vldb.org/2011/files/slides/research30/rSession30-1.pdf>

QSkycube: Efficient Skycube Computation Using Point-Based Space Partitioning

▶ 著者

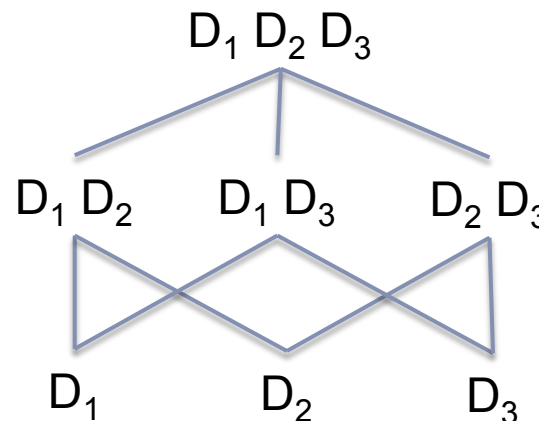
- ▶ Jongwuk Lee, Seung-won Hwang (POSTECH)

▶ 概要

- ▶ 多次元データにおいて, 任意の部分空間に対するskyline (skycube) を効率的に計算する手法を提案.
- ▶ 点ベースの空間分割を利用して効率的に候補を削除.

(例) 3次元データのskycube

Data	D ₁	D ₂	D ₃
a	2	3	2
b	3	5	3
c	5	3	4
d	4	2	3
e	4	3	1
f	6	5	4



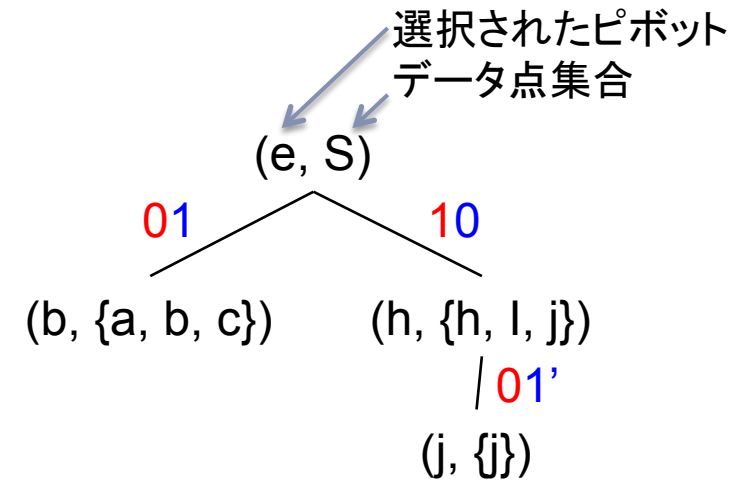
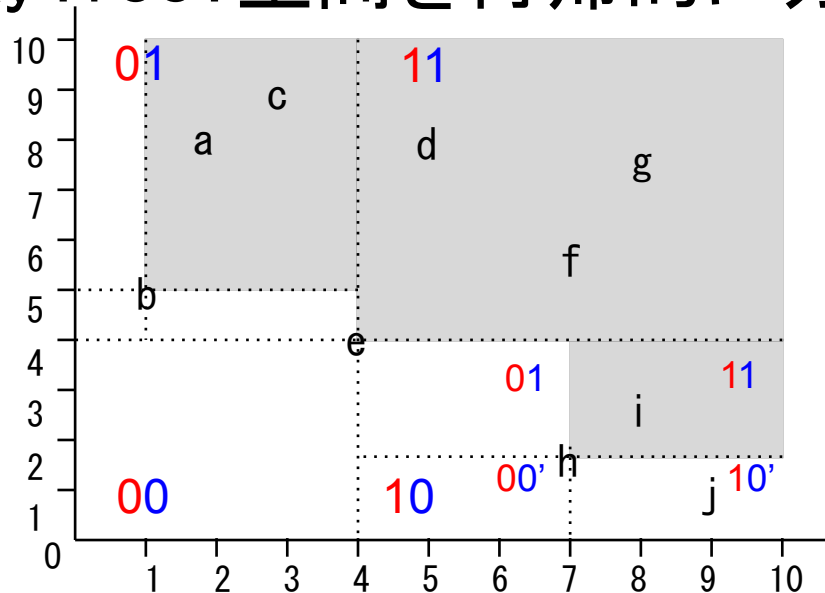
Subspace	Skyline
D ₁	{a}
D ₂	{d}
D ₃	{e}
D ₁ D ₂	{a, d}
D ₁ D ₃	{a, e}
D ₂ D ₃	{a, d, e}
D ₁ D ₂ D ₃	{a, d, e}

Skycube計算のアプローチ

- ▶ 部分空間の計算結果を再利用.
 - ▶ [Y. Yuan, VLDB' 05]
 - ▶ ボトムアップ(BUS)
 - $D_1(D_2)$ のskylineは D_1D_2 のskylineに含まれる.
 - Skyline以外の計算結果は再利用不可能.
 - ▶ トップダウン(TDS)
 - DC法に基づき, 共通する部分空間の構造を利用.
 - 高次元の場合, 構造に基づく最適化が困難.
- ▶ 提案手法
 - ▶ より詳細な計算結果の再利用を検討.
 - ▶ 点ベースの空間分割.

SkyTreeとQSkycube

▶ SkyTree: 空間を再帰的に分割.



▶ QSkycube

▶ SkyTreeを構築し, トップダウンに計算.

▶ 射影

▶ SkyTreeのノード間の支配関係を利用.

□ Vertical / Horizontal

▶ 複数の親の結果を利用(D_iの計算にD₁, D₂, D₃, D₄の結果を利用).

実験

▶ データセット

▶ 合成データ

- ▶ 一様分布 / 相関あり

▶ 比較手法

▶ BUS

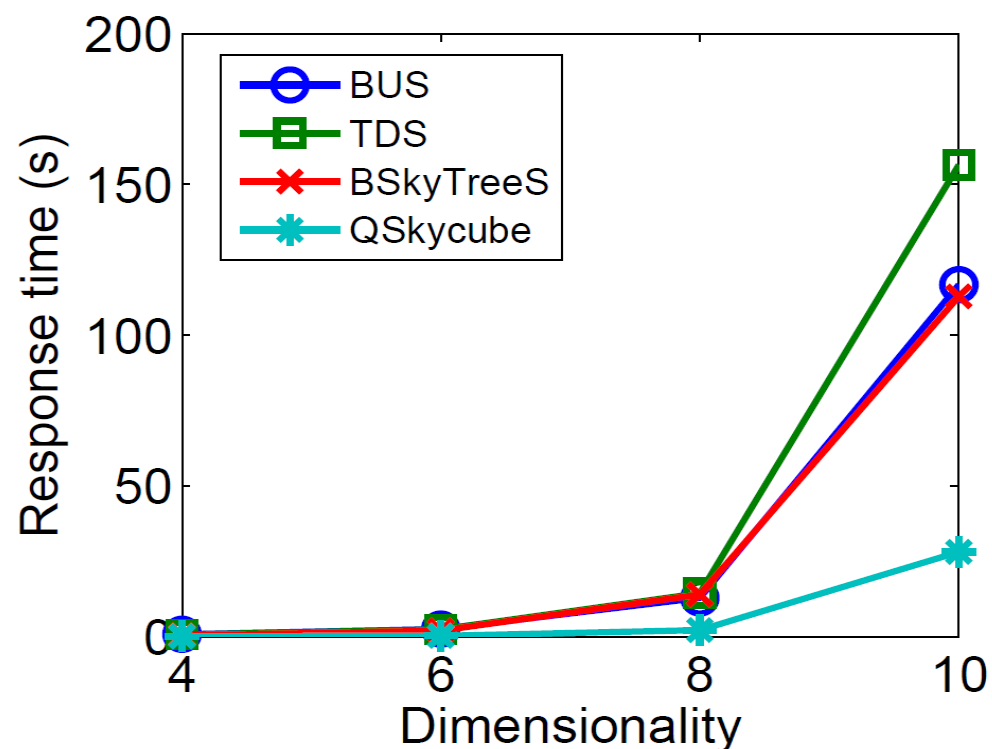
▶ TDS

▶ BSkycubeS

▶ QSkycube

▶ 一様分布

▶ 次元数に対する規模耐性



著者の発表資料より引用:

<http://www.vldb.org/2011/files/slides/research30/rSession30-2.pptx>

A Subsequence Matching with Gaps–Range–Tolerances Framework: A Query–By–Humming Application

▶ 著者

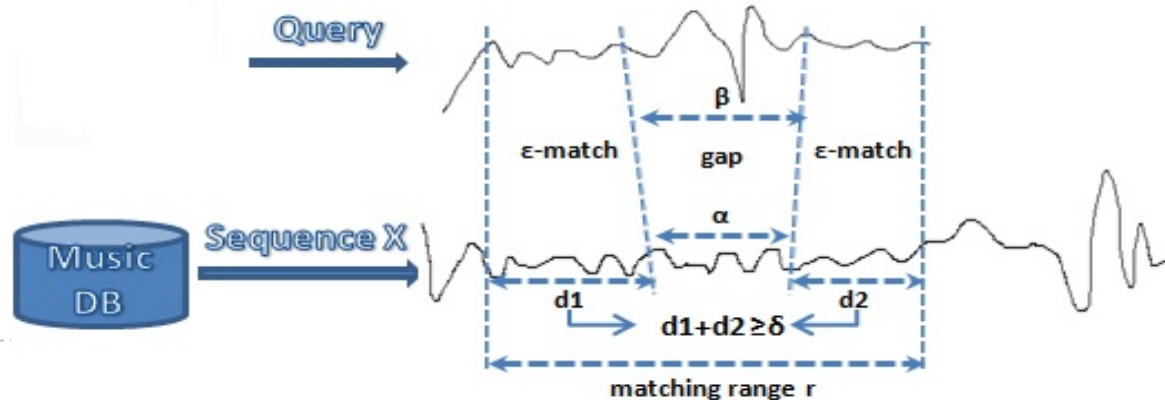
- ▶ A. Kotsifakos (U Athens), P. Papapetrou, J. Hollmen (Aalto U), D. Gunopulos (U Athens)

▶ 概要

- ▶ Query-by-Humming問合せのための部分系列照合手法の提案.
 - ▶ さまざまな制約を考慮
 - 問合せ, ターゲットデータとも, ギャップの存在を許す.
 - 誤差許容範囲を調整可能.
 - マッチする部分系列の最大(最短)長.
- ▶ 新たな類似尺度 SMBGT の提案.

SMBGT: Subsequence Matching with Bounded Gaps and Tolerances

- ▶ シーケンス: 音階 + 持続時間で記述.
- ▶ 問題
 - ▶ 与えられた許容誤差と制約を満たすマッチングする部分シーケンス (Common bounded-gapped subsequence) を探す.
 - ▶ 例: $\alpha = 2, \beta = 1, r = 6, \varepsilon = 1$
 - ▶ $Q = \{6, 3, 10, 5, 3, 2, 9\}, X = \{1, 1, 3, 4, 6, 9, 2, 3, 1\}$
 - $Q = \{6, 3, 10, 5, 3, 2, 9\}$
 - $X = \{1, 1, 3, 4, 6, 9, 2, 3, 1\}$
- ▶ アプローチ
 - ▶ 動的計画法に基づくアルゴリズムを提案.



※元論文より引用.

Approximate Substring Matching over Uncertain Strings

▶ 著者

- ▶ Tingjian Ge, Zheng Li (University of Kentucky)

▶ 概要

- ▶ 不確定な文字列上のマッチングに対する要求が増大.
 - ▶ 例: DNAシーケンサ
- ▶ 不確定文字列上の類似部分文字列マッチング手法の提案.
 - ▶ EED (Expected Edit Distance) [J. Jesters, SIGMOD'10] が, 扱おうとする問題に対して適切でないことを示す.
 - ▶ (k, τ) -マッチング問合せを提案.
 - ▶ (k, τ) -マッチング問合せを効率的に処理するための索引として, 多レベルシグネチャフィルタリングを提案.

定式化・q-gramインデックス

▶ 記号

- ▶ p: パターン文字列
- ▶ X: 不確定文字列集合
 - ▶ 各アルファベットは確率変数(文字レベル(\leftrightarrow)文字列レベル))
- ▶ k, τ : 閾値パラメータ

▶ (k, τ)-マッチング問合せ

- ▶ $\Pr[d(p, X) \leq k] > \tau$ なる X_i の全ての部分文字列を検索.

▶ q-gramインデックス

- ▶ パターンpと文字列の編集距離がk以下かどうかをチェック.
- ▶ パターン p を k+1 個に分割. 編集距離 k 以下なら, q-gramのうち少なくとも一つは完全一致.
- ▶ 残った候補の実際の編集距離をDPで計算し確認.

→ 不確定文字列では, 候補数が増加.

X1: ABRACAD
ABR
BRA
RAC
ACA
CAD

k = 1

p1: BR CC (ED = 1)

p2: BB CC (ED = 2)

多レベルシグネチャフィルタリング

▶ シグネチャ

- ▶ q-gramの前後の文字列からシグネチャ(タグ)を計算.
- ▶ q-gramだけではなく、前後のシグネチャも「ある程度」似ていないと候補になり得ない性質を利用.

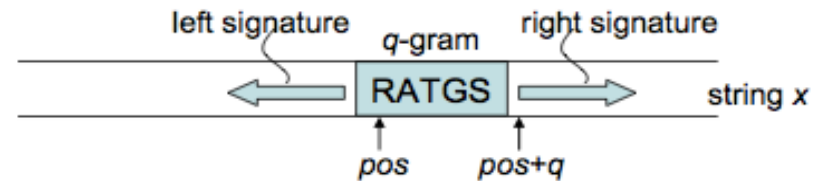


Figure 2. Illustrating the left and right signatures of a string.

▶ 多レベルシグネチャ索引

- ▶ (あるq-gramの)候補位置リストをタグ毎に階層化.

▶ 索引の利用

- ▶ 前後のシグネチャを含めた編集距離を定義.

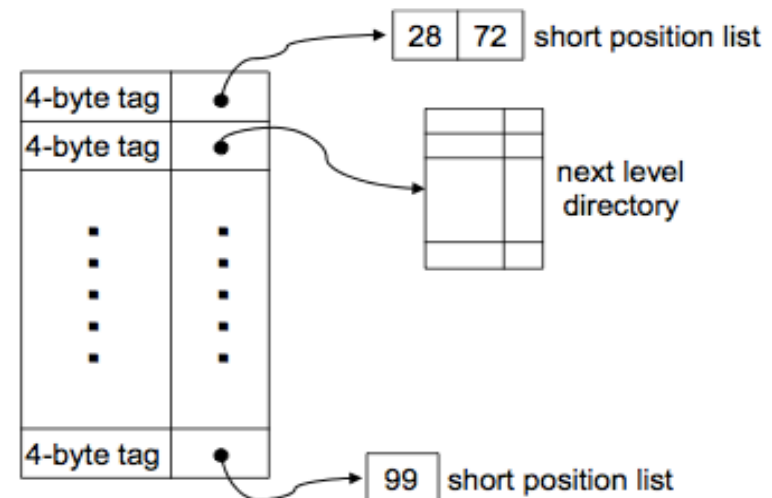


Figure 3. Illustrating multilevel signature filtering structure.

候補の検証アルゴリズム

- ▶ 検証を行う候補を効率的に削減する二つのアルゴリズムを提案.

- ▶ Cumulative Distribution Functions (CDF)

- ▶ 各DPセルにおいて, $k + 1$ 個のペア(確率の上限, 下限)を計算.

$p \setminus x$	C	$G_{.1}A_{.4}T_{.5}$	$G_{.1}A_{.4}T_{.5}$	$G_{.1}A_{.4}T_{.5}$	$G_{.1}A_{.4}T_{.5}$
C	(1, 1)	(0, 0) (1, 1)	(0, 0) (0, 0) (1, 1)		
A	(0, 0) (1, 1)	(.4, .4) (1, 1)	(0, 0) (.64, .64) (1, 1)	(0, 0) (0, 0) (.784, .784)	
T	(0, 0) (0, 0) (1, 1)	(0, 0) (.7, .7) (1, 1)	(.2, .2) (.7, .7) (1, 1)	(0, 0) (.42, .42) (.85, 1)	(0, 0) (0, 0) (.602, .602)

- ▶ Local Perturbation

- ▶ 下限

- adjacentな可能世界を生成し, を摂動.
- $ED \leq k$ となる可能世界の確率の和を計算.

- ▶ 上限

- remoteな可能世界を生成し摂動.
- $ED > k$ となる可能世界の確立の和(p)を計算.
- $1 - p$ が上限を与える.

Figure 6 Computing bounds based on CDF.

評価実験

▶ データセット

- ▶ 実データ: DNA塩基配列, タンパク質アミノ酸配列.
- ▶ 合成データ: DNA, タンパク質の実データから生成.
- ▶ パターンの生成
 - ▶ 長さ15のパターンを三つ抽出.
 - ▶ 六つの不確定文字を含むように修正.

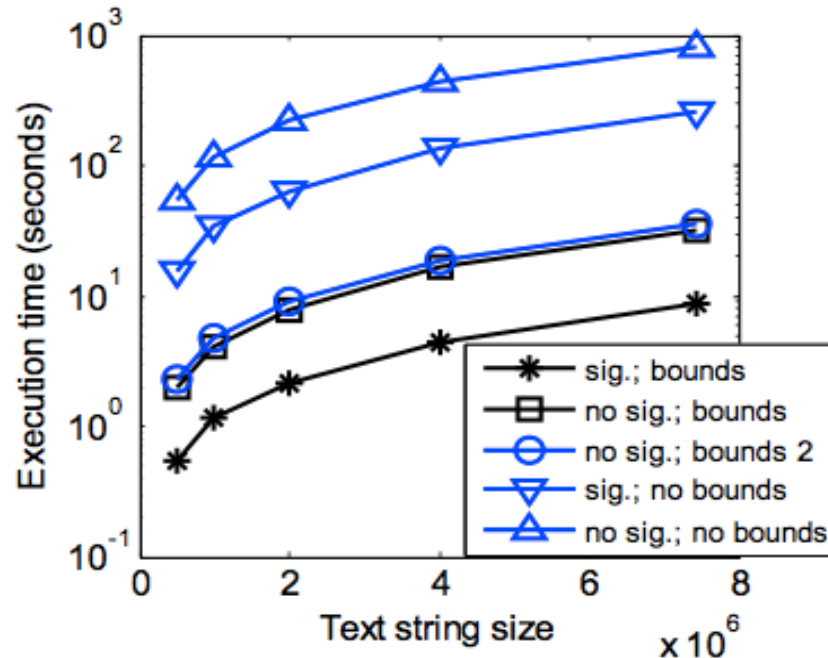


Fig. 7 Running time for various settings.

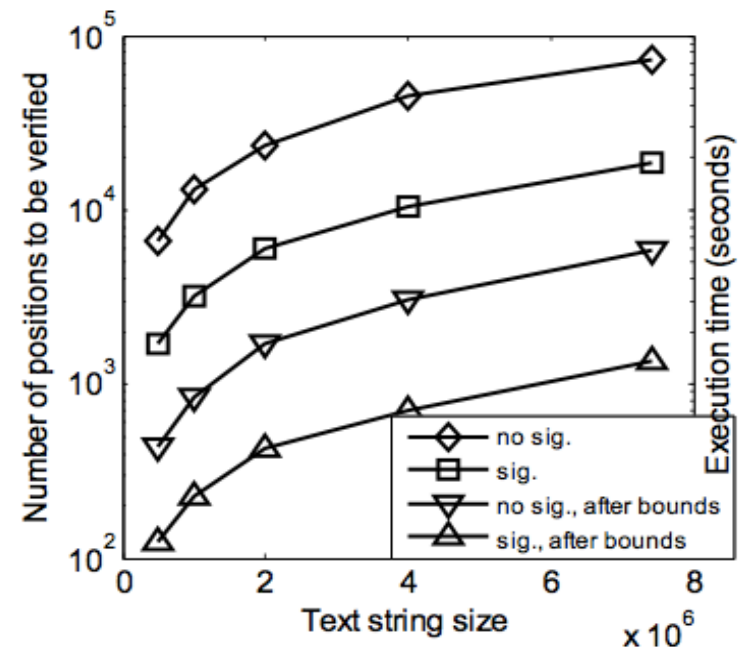


Fig. 11 # of positions to be verified.