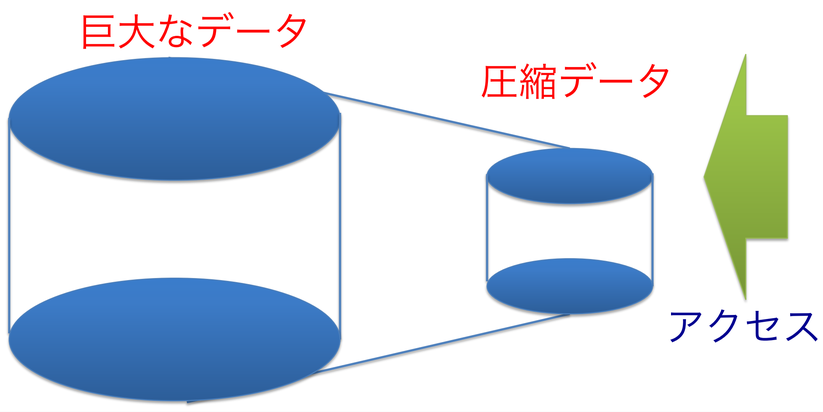
圧縮情報処理ユニット　　田部井　靖生

Succinct Information Processing Unit

Yasuo Tabei

目標：大規模データ解析のため圧縮情報処理技術の開発
利点：計算リソースでの大規模データ解析が可能

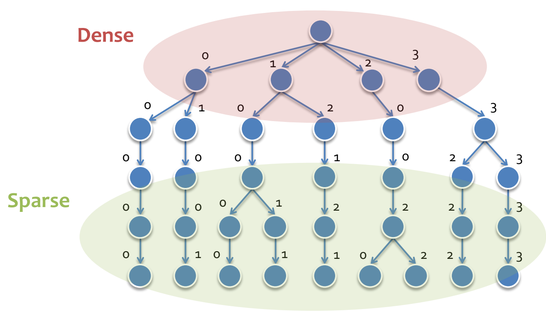


データ解析技術
- 類似度検索
- 線形モデルの学習
- カーネル法
- 文字列処理

etc

整数スケッチ上での大規模類似検索 [IEEE BigData’19]

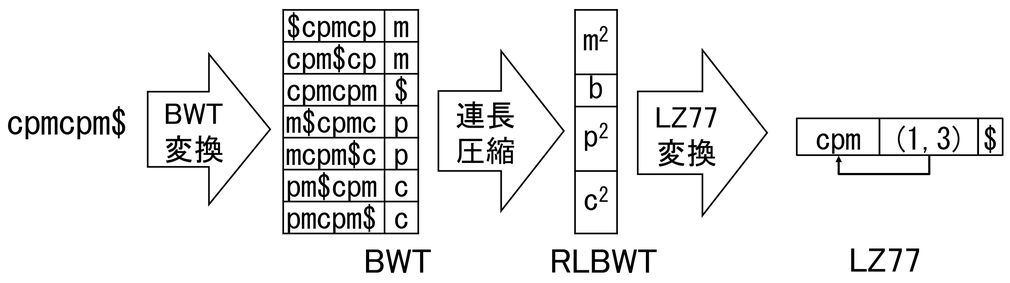
背景：様々なベクトルデータに対する類似検索問題は整数スケッチに対する類似検索問題として効率的に解ける
問題点：既存の検索技術はメモリ効率が悪く大規模データへの適用が困難
提案1：整数スケッチをトライ木で索引付けし高速類似検索を実現
提案2：節点の密度に着目した適応圧縮データ構造を設計
結果（メモリ）：既存法が26GBも要するところ、たった10GBで検索を実現
結果（速度）：同等か、もしくは最大10倍程度高速



メモリ使用量（MB）

		Review	CP	SIFT	GIST
提案	SI-bST	**48**	**1,057**	**9,802**	**1,338**
	MI-bST ($m=2$)	126	3,232	23,159	5,513
	SIH	172	2,329	32,727	4,501
	MIH ($m=2$)	125	4,633	28,876	6,128
	MIH ($m=3$)	160	3,997	26,665	5,744
	HmSearch ($\tau=1,2$)	866	53,097	–	48,456
	HmSearch ($\tau=3,4$)	860	29,396	–	27,337
	HmSearch ($\tau=5$)	860	28,866	–	25,305

RLBWTを用いたより高速なLZ77圧縮 [CPM’19]

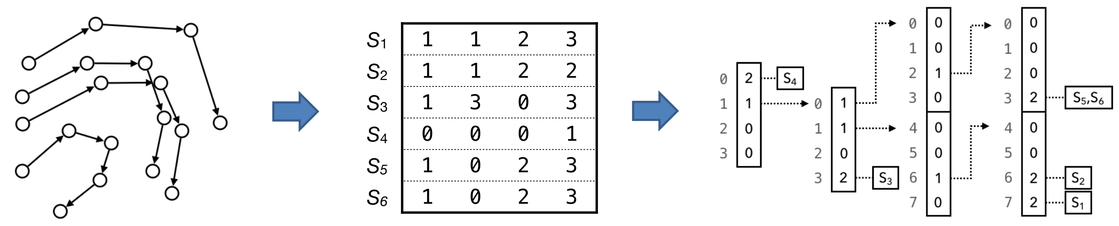


背景：データ圧縮は情報化社会において必須技術
　　現在主流の可逆圧縮アルゴリズムはLZ77圧縮に基づく
問題点：RLBWTを用いたLZ77圧縮 (Policriti and Prezza,’18)
- BWT変換された文字列を連長圧縮（RLBWT）したサイズrに対して$O(r \log n)$の使用メモリー（nは入力文字列長）
- 入力文字列がRLBWTによって小さくなるとき, $r << n$で省領域
- 一方、LZ77フレーズを動的木で探しているので圧縮に$O(n \log r)$時間

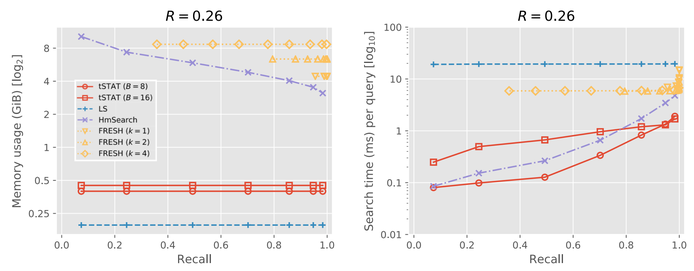
提案：動的木を使わずに静的なデータ構造のみで
　　LZ77フレーズを高速に探す新しいアルゴリズムを提案
結果：使用メモリーを増やすことなく圧縮時間を
　　$O(n \min\{\log\log n, \sqrt{\log r / \log\log r}\})$に改善

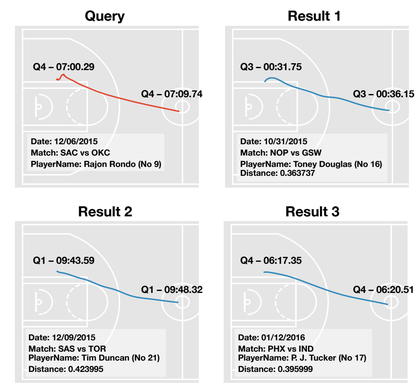
tSTAT：効率的な軌跡類似検索

問題：Fréchet空間上での軌跡類似検索
提案：LSH+マルチインデックス+簡潔トライによる高速かつ省メモリな検索法

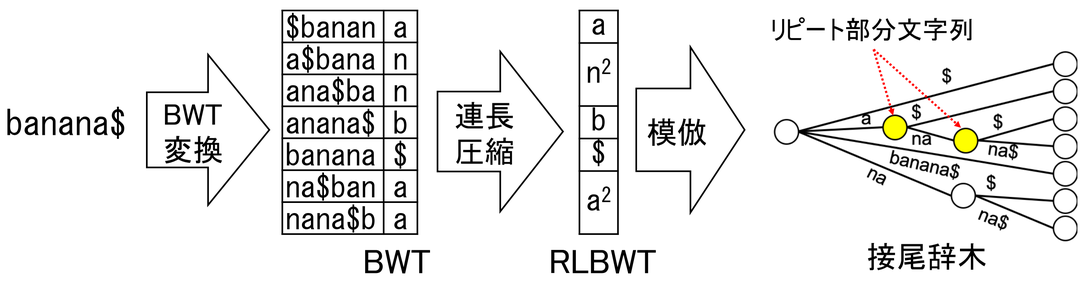


NBA選手の軌跡データを用いた実験結果：
既存法より10倍以上高速かつ省メモリを達成





RLBWT上でのリピート部分文字列発見

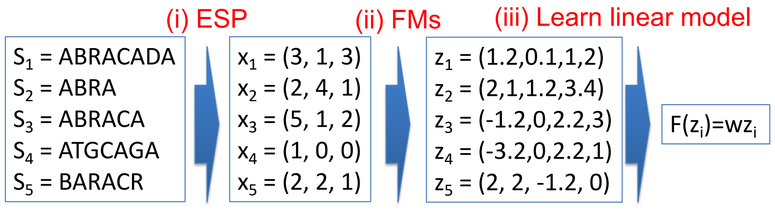


背景：文字列上でのリピート部分文字列の発見は
　　ゲノム配列のモチーフ抽出や文字列上での特徴量抽出として重要
問題点：既存手法の接尾辞木を用いた手法(Okanohara and Tsujii,’09)
- 文字列長nに対して線形時間で発見
- 一方、$O(n \log n)$の使用メモリー

手法：BWT変換された文字列を連長圧縮（RLBWT）し
　　接尾辞木を模倣する新しいデータ構造を提案
結果：RLBWTのサイズrに対して$O(r \log n)$の使用メモリーで
　　リピート発見（$r << n$が成り立つとき省領域）

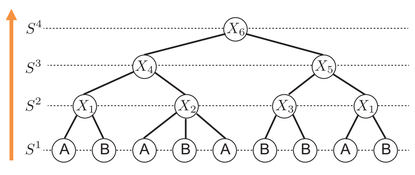
文字列アライメントカーネルのための特徴マップ [ICDM’19]

- 背景：文字列アライメントカーネルは、文字列データ分類、配列相同性が低いDNA/Proteinの分類において有用 [BMCBioinfo.’06]
- 問題点：カーネル法は文字列の長さに対して２乗の計算時間とデータ数の２乗のメモリーが必要
- 手法：Edit-sensitive parsing (ESP)と特徴マップ[Neurips’07]に基づく新しい特徴マップを提案
- 特徴マップに必要なメモリーをO(d×D)からO(d)に削減する手法も提案 (d:入力次元、D:出力次元)
- 結果：大規模テキストの分類が可能になった



Edit-sensitive paring (ESP)
[G.Cormode and S.Muthukrishnan, 2007]

- 概要：文字列Sから構文木を構築するアルゴリズム
- 特徴：同じノードは同じノードラベルを持つ子ノードを持つ
- 文字列Sを整数値特徴ベクトルxを作るのに使うことができる
  - xのそれぞれの要素は対応するノードラベルの出現回数
- メリット：2つの文字列感の編集距離を特徴ベクトル間の編集距離で近似することができる
  (i.e., $EDM(S_i,S_j) \fallingdotseq ||x_i - x_j||_1$)
- 計算時間は文字列の長さに線形時間



A B X_1 X_2 X_3 X_4 X_5 X_6
x = (4, 5, 2, 1, 1, 1, 1, 1)

特徴マップ

$$K(S_i,S_j) = \exp(-EDM(S_i,S_j)/\beta)$$
$$\fallingdotseq \exp(-||x_i - x_j||_1/\beta)) \triangleleft (EDM(S_i,S_j) \fallingdotseq ||x_i - x_j||_1)$$
$$= \int_{R^d} p(\omega) \exp(i(x_i - x_j)\omega)\, d\omega \triangleleft \text{Bochner's theorem}$$
$$= E_{\omega,p}[z_\omega(x_i) z_\omega(x_j)] \quad (z_\omega(x) = \sqrt{2}\cos(x^T\omega))$$
$$= z(x_i)' z(x_j) \quad (z(x) = \sqrt{\frac{1}{d}}(\cos(x^T\omega_1),\ldots,\cos(x^T\omega_D)))$$

ω_i：コーシー分布からサンプリングされる乱数
・O(dD) (d: 入力次元, D:出力次元)メモリーが必要で空間効率が悪い
- Fast food approach [ICML’13]は、RBFカーネルに対する特徴マップを近似するために利用可能
・O(d)メモリーのみ必要な特徴マップを開発した

Experiments

Dataset	Number	#positives	Alphabet size	Average length
Protein	3,238	96	20	607
DNA	3,238	96	4	1,827
Music	10,261	9,022	61	329
Sports	296,337	253,017	63	307
Compound	1,367,074	57,536	44	53

- 5 massive string datasets in real world
- Competitors
  - 5 SVMs with string kernels: LAK [Bioinfo’08], GAK [ICML’11], ESP+Kernel, CGK+Kernel, stk17 [NIPS’17]
  - FMs for alignment kernels: D2KE [KDD’19]
  - SFMEDM: proposed

Training time in second

Method	Protein	DNA	Music	Sports	Compound
SFMEDM(D=128)	5	8	11	204	261
SFMEDM(D=512)	22	34	47	799	1,022
SFMEDM(D=2048)	93	138	193	3,149	4,101
SFMEDM(D=8192)	367	544	729	12,179	16,425
SFMEDM(D=16384)	725	1,081	1,430	24,282	32,651
SFMCGK(D=128)	14	52	26	452	397
SFMCGK(D=512)	60	222	104	1,747	1,570
SFMCGK(D=2048)	237	981	415	7,156	6,252
SFMCGK(D=8192)	969	3,693	1,688	27,790	25,054
SFMCGK(D=16384)	1,937	7,596	3,366	53,482	49,060
D2KE(D=128)	319	4,536	296	8,139	1,641
D2KE(D=512)	1,250	19,359	1244	34,827	6,869
D2KE(D=2048)	5,213	76,937	5,018	140,187	28,116
D2KE(D=8192)	21,208	>48h	19,716	>48h	>48h
D2KE(D=16384)	43,417	>48h	38,799	>48h	>48h
LAK	31,718	-	-	-	-
GAK	25,252	>48h	101,079	>48h	>48h
EDMKernel	20	28	162	>48h	>48h
STK17	3218	917	>48h	>48h	>48h