FY2023/2023年度
Succinct Information Processing Unit
Yasuo Tabei
圧縮情報処理ユニット　田部井 靖生

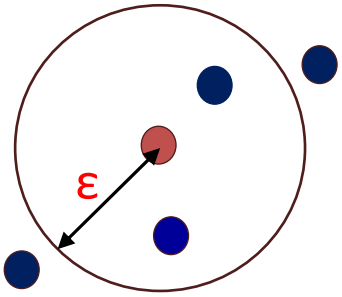
大規模データセットに対する高速かつ省領域な類似度検索技術

ポイント

1. 高次元・大規模データの類似度検索技術
2. 類似データの高速・省メモリ検索
3. ベクトル、グラフ、移動軌跡等のデータ形式対応
4. コサイン、Jaccard、Min-Max等の類似度尺度対応

類似度検索とは

クエリに類似するデータをデータベースから検索する



ε ：閾値

類似度尺度：

類似度	データ
コサイン	画像、テキスト
Jaccard	化合物フィンガープリント
Min-Max	化合物フィンガープリント
ユーグリッド距離	ベクトルデータ
Frechet距離	移動軌跡データ

応用例：
- LLMの質問回答のヒント生成
- テキスト、画像の重複除去
- 移動軌跡データのデータマイニング
- 創薬候補化合物のスクリーニング

手法の概要

入力データ
- ベクトル
- テキスト
- 画像
- 移動軌跡データ
- 化合物

など

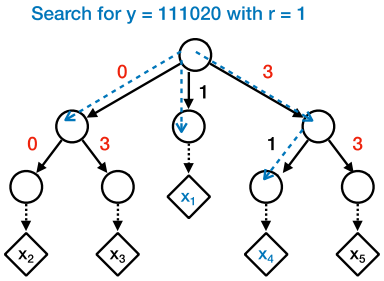
ハッシュ：
- LSH
- LSBC
- CWS

ID　整数列

x_1 : 1 1 3 4 5 6 7 8
x_2 : 0 0 5 6 8 9 2 1
x_3 : 0 3 3 7 5 4 2 1
x_4 : 3 1 6 8 1 3 5 1
x_5 : 3 3 6 8 9 1 2 3

データ間の距離は整数列間のハミング距離として保存

圧縮データ構造（簡潔トライ）

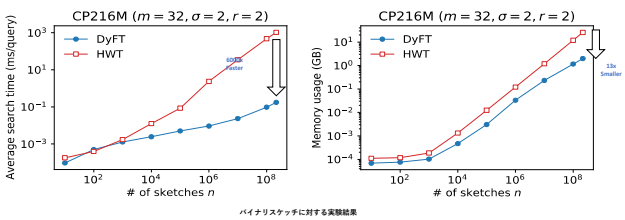


B:111011001100000
A:$x_2x_3x_1x_4x_6$

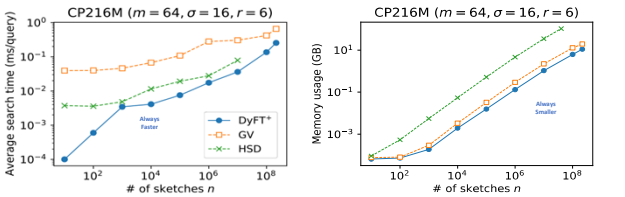
パフォーマンス

実験データ
- 2億の化合物データベクトル
- Liらのb-bit MinHashによりスケッチに変換 [WWW10]

比較手法
HWT [IEEE-TPAMI19] , GV [SIGIR16], HSD [SSDBM13]



バイナリスケッチに対する実験結果



整数スケッチに対する実験結果

公開ソフトウェア

手法	類似度	データ
gWT	コサイン	グラフ
SMBT	Jaccard	0/1ベクトル
bST	Min-Max	ベクトル
Dyft	Min-Max	ベクトル
Frechet_simsearch	Frechet距離	移動軌跡データ

ダウンロードURL

https://x.gd/Iu0H1