Top-k String Similarity Search with Edit-Distance Constraint

Motivation

#1: Data in real world is dirty

Edit Distance: minimum # of single character transformations

*Example:* ED(srajit, seraji) = 2

#2: Hard to define a threshold

#3: Many real applications

- Information retrieval
- Molecular biology
- Bioinformatics
- Data Quality, Data Cleaning

Problem Definition

**Top-k String Similarity Search:** Given a string set S and a query string q, top-k string similarity search returns a string set R ⊆ S such that |R| = k and for any string r ∈ R and s ∈ S − R, ED(r, q) ≤ ED(s, q).

**Definition (Pivotal Entry):**

An entry (i, j) in Ex is called a pivotal entry, if \( D[i + 1][j + 1] > D[i][j] \).

**Definition (Pivotal Triple):**

Given an entry \((n, j)\), one of \(n\)'s children \(n_c\), and a query \(q\), triple \((n, j, n_c)\) is called a pivotal triple, if \( ED(n_c, q[1..j+1]) > ED(n, q[1..j]) \).

**Definition (Pivotal Quadruple):**

A quadruple \([l, u], d, j]\) is a pivotal quadruple, if it satisfies

1. \([l, u]\) is a sub-range of a \(d\)-th level node's range;
2. for any string \(s\) with ID in \([l, u]\), \( ED(s[1..d+1], q[1..j+1]) > ED(s[1..d], q[1..j]) \);
3. strings with ID \(l - 1\) or \(u + 1\) do not satisfy conditions (1) or (2).

Progressive Method: Smallest Cell First

Two Operations: Find Match / Extend

Traditional Method: Dynamic Programming

\[ D_{i,0} = i, D_{0,j} = j, \quad D_{i,j} = \min\{D_{i-1,j} + 1, D_{i,j-1} + 1, D_{i-1,j-1} + 0/1\} \]

Progressive Method: Smallest Cell First

Two Operations: Find Match / Extend

Experiments

implemented in C++, Ubuntu: Intel Xero X5670 2.5GHz CPU and 4 GB memory

- **Datasets**
- **Number of calculated entries**
- **Running Time of Each Methods**
- **Length Distribution**
- **State-of-the-art methods**
- **Scalability**

http://dbgroup.cs.tsinghua.edu.cn/dd/projects/topksearch