Diversity Similarity Join for Big Data

Yas Silva[1], Juan Martinez[1], Pedro Castro[1],
Humberto Razente[2], Maria Nardini[2]
[1]Loyola University Chicago - [2]Univ. Federal de Uberlandia

Yas Silva

Motivation and Contribution

The Problem

◦The Similarity Join can generate massive amounts of result pairs with big datasets

◦Many of the output pairs can be very similar to others adding little value to the analysis process

Our Contribution

◦Distributed Diversity Similarity Join (D2SJ), a distributed operator to diversify the output of the similarity join with big datasets

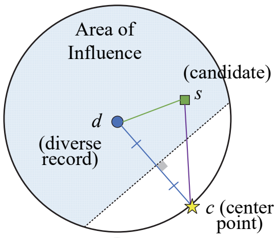
◦Guarantees that each pair is generated once

◦Supports many distance functions and data types

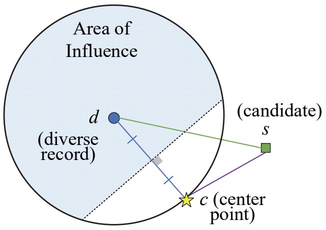
◦Source code of implementation in Apache Spark

Notion of Diversity

s inside of area of influence



s outside of area of influence



Builds on notion introduced by Santos et al. SISAP'15

Evaluation Setup

Algorithms (Spark 3.0)

◦D2SJ, DSJ-CP (direct Spark extension of single-node alg.)

Computer cluster

◦Google Cloud Platform (1 master, 20 workers), node config: 4 vCPUs, 15 GB of memory, 500 GB of disk space

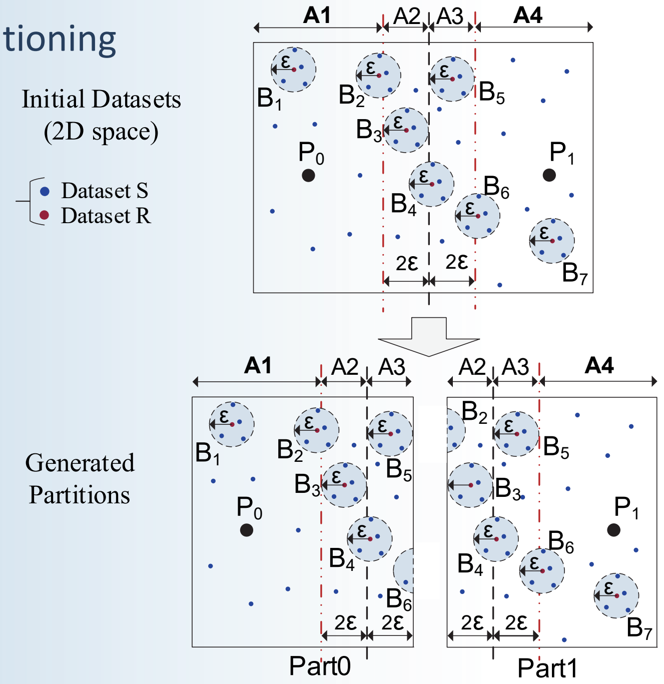
Datasets

◦CoPhIR dataset (16D-282D)

◦Size (SF*N*): *N x* 1M (equally divided between R and S)

◦Ɛ: % of the max potential distance between two records

D2SJ Partitioning



Strategy

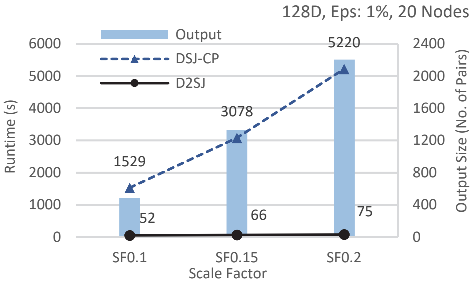
- Partition the input into two partitions such that we can still identify all the *similarity balls* (B_1-B_7) (each ball has all the points in S within **ε** from a point in R)
- Each ball should be finally processed in only one partition producing the diverse pairs in the ball

Solution (using two pivots/partitions)

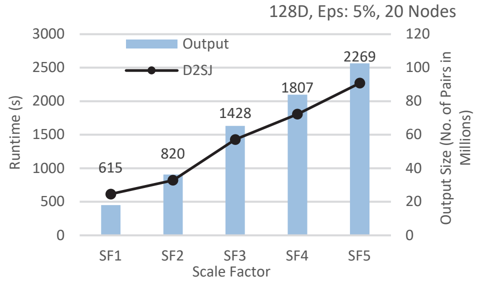
- Partition the input using two pivots (P_0 and P_1) such that each point belongs to the partition of its closest pivot
- Additionally, duplicate the points in the windows regions (A2, A3), generating:
  Part0 = A1+A2+A3, **Part1** = A2+A3+A4
- Each ball is processed in a single partition, the one corresponding to its smallest closest-pivot (using index): Balls B_1, B_2, B_3, and B_4 are processed in Part0 while B_5,B_6, and B_7 in Part1
- Processing a ball (S-points around point r) identifies the subset of diverse pairs (r, s')

Evaluation

Increasing Dataset Size

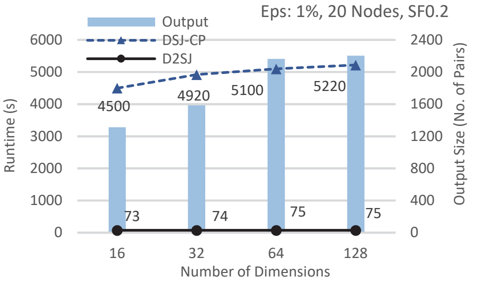


D2SJ vs DSJ-CP

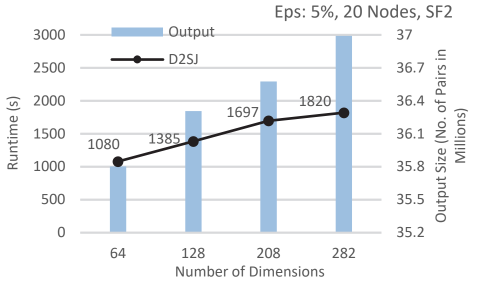


D2SJ with larger datasets

Increasing Dimensionality



D2SJ vs DSJ-CP



D2SJ with higher # of dimensions

This project was supported by an award from the Google Cloud Research program