Approach for the Cluster Engine

Abhishek Pawaskar

Known:

1. No training data to create clusters.
2. No prior information on possibilities of regions of cluster.
3. No prior information on number of clusters.
4. Click Stream is generated for multiple pages.

Preferred Qualities of the solution:

1. Need a solution that dynamically accepts the statistics of the data.
2. Need a solution that can create a cluster on the go.
3. Need a solution that is near real time.
4. Need a solution to handle clustering for multiple pages.

Approach:

1. **Infrastructure**: To add a data holding solution for fast read/writes.
2. **Algorithm**: To create a dynamically changing (plastic) clustering solution.

[Here the Ideology is to divide the Data Layer (Redis) & Compute Layer (cluster-engine) of the Clustering Model]

1. **Infrastructure**:
   Due to fast read/write speeds & ease of management, Redis is chosen here.

2. **Algorithm**: (Refer Diagram in next page for complete view)

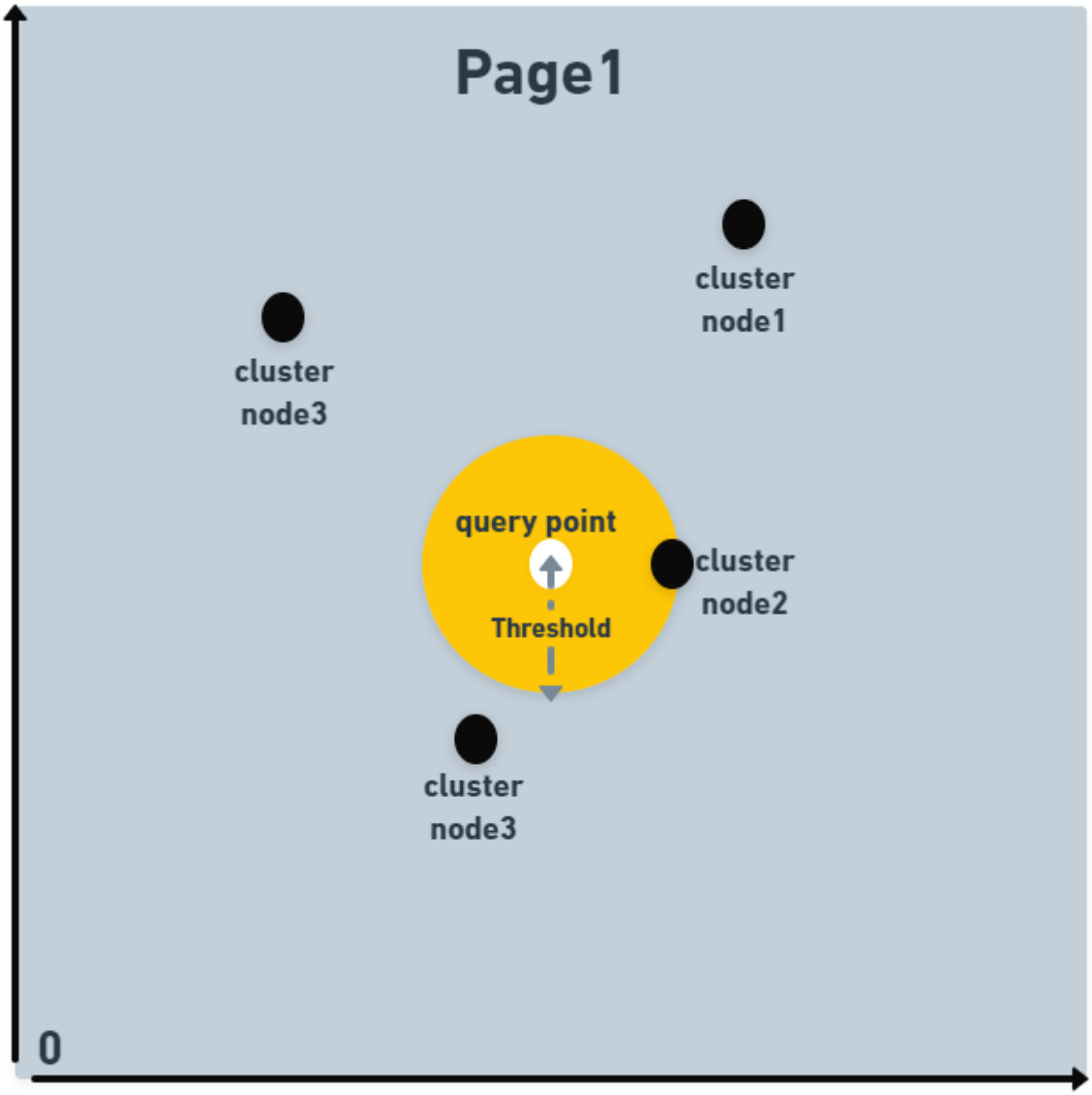
   Predefined aspects of this algorithm:
   a. **Cluster node:** A point in space that acts as a cluster head for that cluster and is an also a part of the click stream record in the Data Base.
   b. **Threshold:** A distance value that determines whether the query point is qualified to be considered in the calculation of cluster determination with the cluster nodes.
   c. **Cluster node points:** X & Y coordinated of that Cluster node.
   d. **Cluster_node_list (redis):** List of all the cluster nodes for that page.
   e. **Cluster_node_points(redis):** List of X & Y coordinated of that Cluster node.

- **Algorithm Flow for Save & Predict Cluster** (with Data Operations):
  1. Accept the request and do data validation.
  2. Create a record in the Table (PostgreSQL)
  3. Check with the cluster node list if there exist any nodes for that page (Redis)
     a) If No, then create a new cluster with the same given coordinates as the cluster node for that page. (Redis)
     b) If Yes, then pull up all the cluster nodes of that page for calculation. (Redis)
        i. Process the coordinates of points in array format
        ii. Compute distance between query point and cluster heads (cluster nodes)
        iii. Use Threshold to bring down the additional compute load (to choose top nodes)
        - If no cluster node is qualified, then created a new cluster in page. (Redis)
        - Else, choose the best cluster node and generated the cluster id as result.

- **Algorithm Flow for Predict Cluster** (with Data Operations):

  Step 3 from Algorithm Flow for Save & Predict Cluster

Diagram To explain the Algorithm Execution with Page specific Cluster heads



Advantages of this approach:

1. Query result updates and can become more accurate as the engine discovers more data.
2. No dependency on number of K-clusters to be fixed.

Disadvantages of this approach:

1. This method relies heavily on the hyperparameter (Threshold)