Horizontal Fragmentation Based on data Semantics

UNDER SUPERVISION OF

PROF. ANANTH NARAYAN V.S.

PRESENTED BY

ABHISHEK KUMAR

202IT001

Problem

- There are many existing fragmentation method already available to us but the problem with these methods are that they can only be implemented when list of query is available.
- Also most of the existing horizontal fragmentation method do not consider the data relationship for the fragmentation, but user usually query related data.
- So, In this paper they have considered the data semantics or relationship for the fragmentation.
- Now for putting the related data together clustering technique will be used and after applying the clustering algorithm we will get different sets/clusters of related data.
- Now, after getting the clusters of related data we can divide our database into various fragments.

Methodology

Firstly we find the clusters of related data.

Do fragmentation based on clustered data.

Find the Jaccard similarity for each pair of the fragment.

Merge the most similar fragments until we get the equal number of fragments as the number of nodes.

Step 1: Methodology

- For making cluster of similar data, **K-prototype algorithm** has been used.
- K-prototype algorithm is suitable for the mixed data (means we can have numeric, categorical data together).

K-Prototype algorithm

- First select k random points.
- Now instead of using Euclidian distance for finding the similarity , We shall use Gower's similarity co-efficient.
- Gower's similarity co-efficient = number of similar valued between 2 data points.
- Now, (data, centroid) having more Gower's similarity co-efficient will be kept together.
- Calculate the new centroid using modes instead of mean.
- Repeat these step until previous and current cluster's would be same.

Step 2: Methodology

- After getting the clusters assign corresponding cluster number to each data point.
- Based on cluster number apply primary horizontal fragmentation.

Step 3 : Methodology

- Jaccard similarity defined as :

  J(A,B)=$(A \cap B)/(A \cup B)$ where A and B are the fragments.
- Using this equation find the Jaccard similarity of each pair.

Step 4 : Methodology

- Join/merge the most similar fragments until **# of fragments greater than the # of nodes.**

Environment details and Technology used

Environment details

1. Google cloud platform for creating distributed environment.
2. Postgres SQL database
3. PgAdmin used for client
4. Jupyter notebook for writing code

Technology Used

1. Python
2. SQL
3. Used libraries : numpy, pandas Kmodes, time, sqlalchemy,psycopg2

Stepwise Implementation details

1. First load the dataset(Banking dataset from UCI KDD database repository)
2. After loading do pre-processing on the data.
3. Apply K-prototype algorithm and find the clusters of related data.
4. After getting the cluster assign corresponding cluster number to each data.
5. Then apply primary horizontal fragmentation based on cluster number.
6. After getting the fragmentation apply merging if #of nodes< # of fragments.
7. Create the nodes on google cloud and assign a static IP to each node.
8. After getting the fragments , insert these fragment into suitable locations using sqlalchamy library.
9. Run the queries and compare the result.

Results

Query	Time wds	Time ds
Tuples having Age=39	1358ms	2ms
Tuples having Age=20	4ms	720
Tuples having Age=29	526	0

My Contribution

- Do a comparative analysis with different clustering algorithm for finding the relationship between data's.
- Check on which type of data which clustering algorithm is suitable so that query would processed faster.

Thank You