Similarity Group-by (SGB)

- We propose to extend the standard grouping operators to group similar or approximate values.
- The main goal of SGB is to generate more meaningful and useful similarity-based groupings than those of the regular group-by while maintaining:
  - Low running time
  - Good scalability properties
  - Efficient integration with the query processing engine

Optimization

- Materialized views can be used to answer similarity queries (details in paper)
- Eager/Lazy similarity aggregation (Main theorem)

Implementation (PostgreSQL)

- The Parser
  - Extended the grammar rules and parse tree structure
- The Planner/Optimizer
  - Made use of the RHS input plan tree of aggregation nodes
  - Each aggregation node processes 1 SGA and 1 or more GAs
  - SGAs can be ordered to reduce number of flowing tuples
- The executor
  - Hash-based approach
  - Single plane sweep approach

Performance Evaluation (TPC-H)

- Performance while increasing dataset size
- Performance of generating similarity groups with GB vs. SGB
- Performance while increasing SGAs
- Performance of complex queries

Use Case Scenarios

- Grouping features are extensively used in OLTP, OLAP, and decision support systems.
- Many applications scenarios can benefit from queries that take advantage of similarities in the data (biology, sensor networks, business)

Sensor Network: Track surrounding phenomena

- \( \text{SensorsReadings} \)
- \( \text{Temperature} \)
- \( \text{Pressure} \)
- \( \text{Thresholds} \)
- \( \text{MAXIMUM_ELEMENT_SEPARATION} \)
- \( \text{MAXIMUM_ELEMENT_SEPARATION} \)
- \( \text{MAXIMUM_GROUP_DIAMETER} \)
- \( \text{MAXIMUM_GROUP_SEPARATION} \)

SGB: Three Instances

Unsupervised SGB

- \( \text{SELECT select_expr,} \) \( \text{...} \) \( \text{FROM table_references} \) \( \text{WHERE where_condition} \)
- \( \text{GROUP BY a1 AROUND (T1), a2 AROUND (T2)} \)
- \( \text{GROUP BY a1 DELIMITED BY (T1), a2 DELIMITED BY (T2)} \)
- \( \text{SELECT AVG(Temperature) AS SS1, CNT} \)
- \( \text{GROUP BY a1 DELIMITED BY (T1), a2 DELIMITED BY (T2)} \)

Supervised Similarity Group Around

- \( \text{SELECT select_expr,} \) \( \text{...} \) \( \text{FROM SensorsReadings} \) \( \text{WHERE where_condition} \)
- \( \text{GROUP BY a1 AROUND (T1), a2 AROUND (T2)} \)
- \( \text{GROUP BY a1 DELIMITED BY (T1), a2 DELIMITED BY (T2)} \)
- \( \text{SELECT G1 on Seg1, G2 on Seg2} \)
- \( \text{SELECT L.l_discount as DcntLevel, O.o_clerkType, sum(R1.CNT)} \)
- \( \text{GROUP BY O.o_clerkType, sum(R1.CNT)} \)

Supervised Similarity Group with Delimiters

- \( \text{SELECT select_expr,} \) \( \text{...} \) \( \text{FROM SensorsReadings} \) \( \text{WHERE where_condition} \)
- \( \text{GROUP BY a1 AROUND (T1), a2 AROUND (T2)} \)
- \( \text{GROUP BY a1 DELIMITED BY (T1), a2 DELIMITED BY (T2)} \)
- \( \text{SELECT \text{parse}}(\text{value}, \text{threshold}) \) \( \text{AS SS1, CNT} \)
- \( \text{GROUP BY a1 DELIMITED BY (T1), a2 DELIMITED BY (T2)} \)

Business question: Study the buying behavior of clients that order by their company.

- Groups all orders according to the buying behavior of clients.
- Each order is assigned to the group of the closest average order.