Additional information

ACM Reference Format:
. 2024. Additional information. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Criteria for categories of datasets analysis aspects

The categorization of the results for each aspect is made based on the following criteria:

1.1 SQL queries

- Structural variety
  - Low: At least 50% of the dataset contains queries of type "select-from" or "select-from-where".
  - Medium: Structural combinations that contain at least one of the order by, group by having, or limit clauses exist in a percentage higher than 50%.
  - High: At least 40% of the queries contain nestings and the average depth in these queries is more than 2.
- Operator variety
  - Low: Combinations of only joins, aggregates, logical and comparison operators comprise more than 75% of the dataset.
  - Medium: Combinations containing any of the rest of the operator types constitute more than 25% of the dataset.
  - High: Combinations containing any of the rest of the operator types constitute more than 50% of the dataset.
- Operator usage
  - Low: The average number of operators is smaller than 5.
  - Medium: The average number of operators is between 5 and 15.
  - High: The average number of operators is more than 15.
- Schema usage
  - Low: Less than 5 columns or 3 tables are used in at least 75% of the queries in the dataset.
  - Medium: 5-10 columns or 3-5 tables are used in more than 25% of the dataset.
  - High: More than 10 columns or more than 5 tables are used in more than 25% dataset.
- Content usage
  - At most 2 values are used in at least 50% of the dataset.
  - 3-10 values are used in at least 50% of the dataset.
  - More than 10 values are used in at least 25% of the dataset.

1.2 Natural language questions

- Schema linkage
  - Low: The average percentage of exact schema references is below 20%.
  - Medium: The average percentage of exact schema references is between 20% - 50%.
  - High: The average percentage of exact schema references is higher than 50%.
- Lexical Complexity
  - Low: avg rarity =< 0.2 or avg lexical density =< 0.4
  - Medium: 0.2 < avg rarity =< 0.5 and 0.4 < avg lexical density =< 0.6
  - High: avg rarity > 0.5 or avg lexical density > 0.6
- Syntactic Complexity
  - Low: avg dependencies depth <=3 or avg length <=7
  - Medium: 3 < avg dependencies depth =< 5 and 7< avg length =< 13
  - High: avg dependencies depth > 5 or avg length > 12
- Readability
  - Low: The average readability score is below 60.
  - Medium: The average readability score is between 60 and 80.
  - High: The average readability score is more than 80.

1.3 Databases

- Schema complexity
  - Low: The average number of tables in the databases is less or equal to 5 or the average number of columns in the databases is less than 100.
  - Medium: The average number of tables in the databases is between 6 and 25 or the average number of columns in the databases is between 100 and 1000.
  - High: The average number of tables in the databases is more than 25 or the average number of columns in the databases is more than 1000.
- Schema quality
  - Low: There are databases with a percentage of explainable schema elements less than 50.

Conference'17, July 2017, Washington, DC, USA
© 2024 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

  - Medium: The percentage of explainable schema elements is more than 50 in all databases.
  - High: The percentage of explainable schema elements is more than 80 in all databases.
- Database size
  - Low: There are databases that contain less than 1000 total rows.
  - Medium: All the databases contain total rows between 1k and 1m.
  - High: There are databases containing more than 1 million rows.