DELHI TECHNOLOGICAL UNIVERSITY

DATABASE MANAGEMENT SYSTEM (MC-302)

Midterm Innovative Project

Research Paper Analysis:

Similarity Metrics for SQL Query Clustering

G. Kul, D. T. A. Luong, T. Xie, V. Chandola, O. Kennedy and S. Upadhyaya

IEEE Transactions On Knowledge And Data Engineering, December 2018

SUBMITTED BY:

Aiman Siddiqua - 2K18/MC/008

Apoorva - 2K18/MC/019

SUBMITTED TO:

Goonjan Jain

1 ABSTRACT

- Database access logs are the starting point for many forms of database administration, from database performance tuning, to security auditing, to benchmark design, etc.
- Clustering is the first step towards understanding the massive query logs, for an analyst to extract broad patterns.
- Most clustering methods use pairwise similarity, which can be difficult for SQL Queries, especially without underlying database schema or data.
- In the paper, query similarity is computed relying only on the query structure.
- Three query similarity heuristics were applied to query clustering.
- To improve the accuracy of the heuristic, a generic feature engineering method was applied using classical query rewrites to standardize query structure.
- The proposed strategy results in a significant improvement in the performance of all three similarity heuristics.

2 INTRODUCTION

Query clustering has different motives. Some methods prefer using the log as a resource to collect information to build user profiles, and the others utilize structural similarity to perform tasks like query recommendation, performance optimization, session identification and workload analysis.

The three metrics were evaluated on two types of data: Human-authored and Machine-generated. Three evaluation data sets were used:

- A large set of student authored queries released by IIT Bombay.
- A smaller set of student queries gathered at the University at Buffalo, and released as part of this publication.
- SQL logs that capture all activities on 11 Android phones for a period of one month.

The paper accomplishes the following, A survey of existing SQL query similarity metrics, an evaluation of these metrics on multiple query logs, and applying query standardization.

Summary of UB Exam Dataset

Year	2014	2015
Total number of queries	117	60
Number of syntactically correct queries	110	51
Number of distinct query strings	110	51
Number of queries with score > 50%	62	40

Summary of PocketData Dataset and Google+

	Pocket Dataset	Google+
All queries	45,090,798	2,340,625
SELECT queries	33,470,310	1,352,202
Distinct query strings	34,977	135

Summary of IIT Bombay Dataset

Question	Total number of queries	Number of parsable queries	Number of distinct query strings
1	55	54	4
2	57	57	10
3	71	71	66
4	78	78	51
5	72	72	67
6	61	61	11
7	77	66	61
8	79	73	64
9	80	77	70
10	74	74	52
11	69	69	31
12	70	60	22
13	72	70	68
14	67	52	52

3 LITERATURE REVIEW

Paper title	Motivation	Features	Distance Function
Agrawal et al. (2006)	Q. reply importance	Schema, rules	Cosine similarity
Giacometti et al. (2009)	Q. recommendation	Difference pairs	Difference query
Yang et al. (2009)	Q. recommendation	Selection/join, projection	Jaccard coefficient on the graph edges
Chatzopoulou et al. (2011)	Q. recommendation	Syntactic element frequency	Jaccard coefficient and cosine similarity
Aouiche et al. (2006)	View selection	Selection/join, group-by	Hamming distance
Aligon et al. (2014)	Session similarity	Selection/join, projection, group-by	Jaccard coefficient
Makiyama et al. (2016)	Workload analysis	Term frequency of projection, selection/join, from, group-by and order-by	Cosine similarity

4 FEATURE ENGINEERING

Clustering quality was significantly improved by standardizing certain SQL features into a more regular form with various techniques. This is known as regularisation. It aims to produce a new query that is more likely to be structurally similar to other semantically similar queries.

4.1 REGULARIZATION RULES:

- **Canonicalize Names and Aliases:** First regularization step attempts to create a canonical naming scheme for both attributes and tables.

- **Syntax Desugaring:** SQL's redundant syntactic sugar was removed by following basic pattern-replacements as shown in the following table:

Syntactic Desugaring

Before	After
$b \{>, \geq\} a$	$a \{<, \leq\} b$
x **BETWEEN** (a,b)	$a \leq x$ **AND** $x \leq b$
x **IN** $(a, b, \ldots)$	x=a **OR** x=b **OR** ...
isnull(x,y)	**CASE WHEN** x is null **THEN** y **END**

- **EXISTS Standardization:** Although SQL admits four classes of nested query predicates: (EXISTS, IN, ANY, and ALL), the EXISTS predicate is general enough to capture the semantics of the remaining operators. Queries using the others are rewritten:

```
x IN (SELECT y ... ) becomes EXISTS (SELECT * ... WHERE x=y)
x < ANY (SELECT y...) becomes EXISTS (SELECT *... WHERE x< y)
x < ALL (SELECT y...) becomes NOT EXISTS (SELECT *... WHERE x≥ y)
```

- **DNF Normalization:** We normalize all boolean-valued expressions by converting them to disjunctive normal form(DNF).

- **Commutative Operator Ordering:** We standardize the order of expressions involving commutative and associative operators (e.g., ^,v , + , and *) by defining a canonical order of all operands and traversing the expression tree bottom-up to ensure consistent order of all operands.

- **Nested Query De-correlation:** A common database optimization called nested-query de-correlation converts some EXISTS predicates into joins for more efficient evaluation.

- **OR-UNION Transform:** We use a regularization transformation that exploits the relationship between OR and UNION.

- **Union Pull-Out:** Since the prior transformation may introduce UNION operator in nested subqueries, we push selection predicates down into the union as well.

5 EXPERIMENTAL METHODS

5.1 CLUSTERING VALIDATION MEASURES

Clustering validation measures are used to validate the quality of a labeled dataset by estimating two quantities:

1. the degree of tightness of observations in the same label group and
2. the degree of separations between observations in different label groups.

5.1.1 SILHOUETTE COEFFICIENT

For every data point in the dataset, its silhouette coefficient is a measure of how similar it is to its own cluster in comparison to other clusters.
In particular, the silhouette coefficient for a data point i is measured as

$$\frac{b(i) - a(i)}{max(a(i), b(i))}$$

where a(i) is the average distance from i to all other data points in the same cluster and b(i) is the average distance from i to all other data points in the closest neighboring cluster.

5.1.2 BETACV MEASURE

The BetaCV measure is the ratio of the total mean of intra-cluster distance to the total mean of inter-cluster distance. The smaller the value of BetaCV, the better the similarity metric characterizes the cluster partition of queries on average.

5.1.3 DUNN INDEX

The Dunn Index is defined as the ratio between minimum distance between query pairs from different clusters and the maximum distance between query pairs from the same cluster. Higher values of Dunn Index indicate better worst-case performance of the clustering metric.

5.2 SIMILARITY METRICS

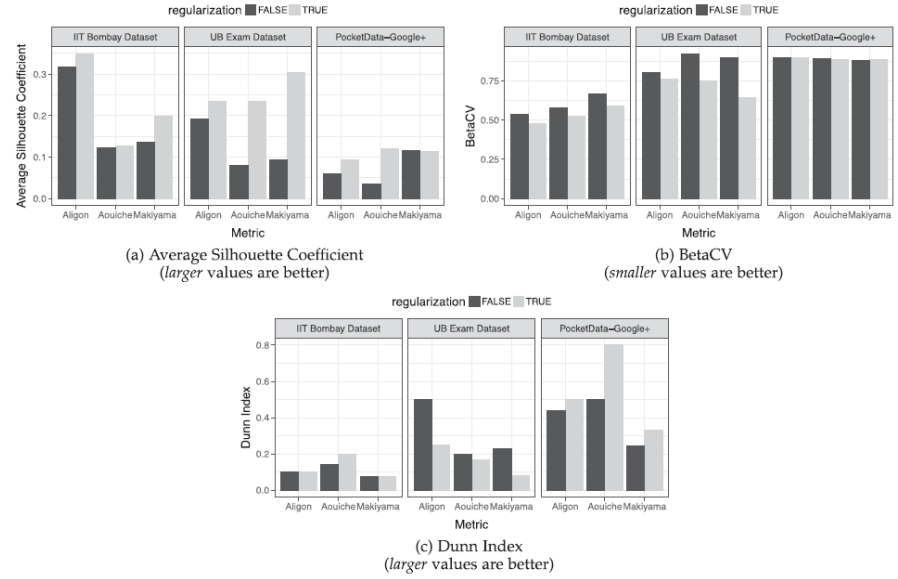
We use the similarity metrics as defined in Aouiche et al. (2006), Aligon et al. (2014) and Makiyama et al. (2016).

6 RESULTS

The aim of the first experiment was to evaluate which similarity metric can best capture the task performed by each query. The black columns in Fig. 1 show a comparison of three similarity metrics using each of the three quality measures. As can be seen in Fig. 1, Aligon seems to work the best for the workloads.

Next experiment evaluated the effectiveness of regularization by applying it to each of the three metrics. Fig. 1 showed the values of three validation measures for each of the three similarity metrics, both with and without regularization.

Finally, the Silhouette Coefficient for Aligon Metric was evaluated.



(a) Average Silhouette Coefficient (*larger* values are better)

(b) BetaCV (*smaller* values are better)

(c) Dunn Index (*larger* values are better)

Fig. 1. Clustering validation measures for each metric with and without regularization step.

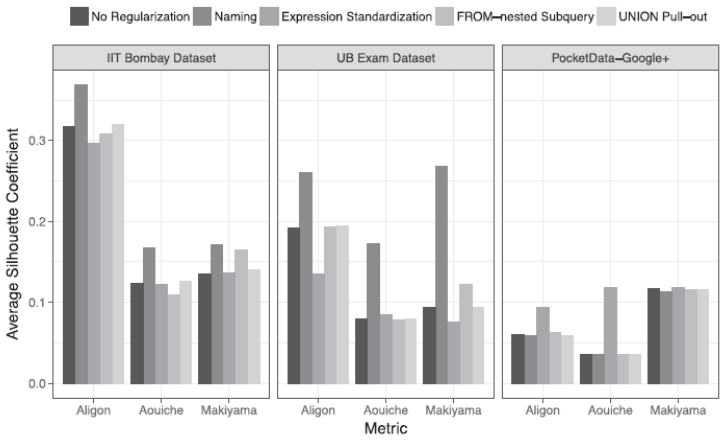


Fig. 3. Effect of each module in regularization.

From the graphs, the overall effects of regularization could be seen.

The conclusion was that different workloads have different characteristics and no one similarity metric surveyed was always good. The feature engineering steps provided an improvement across the board because they addressed the error reasons that were identified.