Qualitative Analysis of the SQLShare Workload for Session Segmentation

Verónika Peralta, Willeme Verdeaux, Yann Raimont, Patrick Marcel

LIFAT – University of Tours – France

DOLAP – Lisboa – March 2019

Interactive Data Exploration

```
SELECT year(Orderdate) AS Year, Nation,
       Mfgr, sum(Quantity) AS Qty
FROM LineOrder NATURAL JOIN Customer
     NATURAL JOIN Part
WHERE
  AN
GROU
```

Year	Nation	Mfgr	Qty
2016	Argentina	MFGR#1	130
2016	Argentina	MFGR#2	235
2016	Argentina	MFGR#3	35
2016	Argentina	MFGR#4	130
…	…	…	…

```
SELECT year(Orderdate) AS Year, Nation,
       sum(Quantity) AS Qty
FROM LineOrder NATURAL JOIN Customer
     NATURAL JOIN Part
WHERE Mfgr = 'MF
  AND Nation = '
GROUP BY year(Or
```

Year	Nation	Qty
2016	Argentina	35
2015	Argentina	200
2014	Argentina	190
2013	Argentina	175
…	…	…

```
SELECT year(Orderdate) AS Year, Nation,
       sum(Quantity) AS Qty
FROM LineOrder NATURAL JOIN Customer
     NATURAL JOIN Part
WHERE M
  AND (N
  O
GROUP B
```

Year	Nation	Qty
2016	Argentina	35
2016	Brazil	240
2015	Argentina	200
2015	Brazil	210
…	…	…

```
SELECT year(Orderdate) AS Year, City,
       sum(Quantity) AS Qty
FROM LineOrder NATURAL JOIN Customer
     NATURAL JOIN Part
WHERE
  AN
GROU
```

Year	City	Qty
2016	Buenos Aires	0
2016	La Plata	12
2016	Mendoza	15
2016	Rosario	8
…	…	…

- **Discovering explorations is a first step for:**
  - Discovering user intents, focus/explorative zones, user expertise…
  - Adapted visualization, recommendation of data/queries, personalization…
- **Past experience with OLAP queries**

SQLShare

- **Multi-Year SQL-as-a-Service experiment** [Jain et al. 2016]
  - A large log of hand-written SQL queries over user-uploaded datasets.
  - Limitations:
    - Not all datasets are available
    - No timestamps
    - No ground truth

11 137 SQL statements
(10 668 SELECT statements)
57 users
3336 user's datasets

- **Other SQL workloads**
  - SDSS workload mixes hand-written and bots' queries [Singh et al. 2006]
  - Other workloads are too small

Challenge: Session segmentation

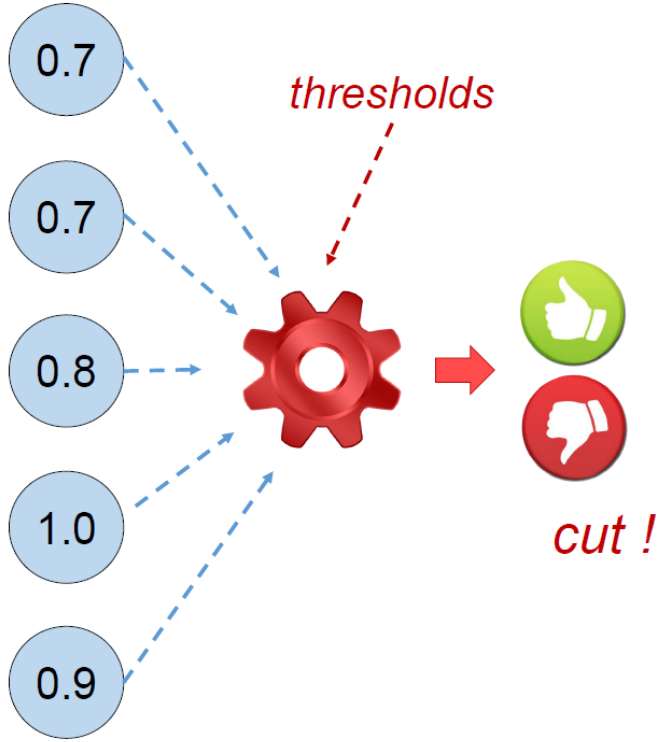
- **Detecting begin-end of explorations in a log of user sessions**

- **Previous work**
  - In web community: cut after 30 minutes of inactivity
  - In SDSS workload: same 30-minutes delay

- **Simple but not reliable**
  - And timestamps not always available

Approach: Similarity-based Segmentation

- **Our approach: cut when queries are dissimilar enough**
  - Several similarity indexes based on query features
    - Projections, selections, aggregations and tables
  - Voting strategy
- **Extensible to other features and indexes**



Feature extraction

Query text

```
SELECT latBin,
    longBin,
    COUNT(species)
FROM [690].[All3col]
WHERE latBin > 0
GROUP BY latBin,longBin
HAVING COUNT(species) > 5;
```

- **Query parts**
  - Projections
    - latBin
    - longBin
    - COUNT(species)
  - Selections
    - latBin > 0
    - COUNT(species) > 5
  - Aggregations =
    - COUNT(species)
  - Tables
    - [690].[All3col]

- **Intrinsic metrics**
  - Nb Projections
  - Nb Selections
  - Nb Aggregations
  - Nb Tables
- **Relative metrics**
  - Nb common Projections
  - Nb common Selections
  - Nb common Aggregations
  - Nb common Tables
  - Relative Edit Distance
  - Jaccard Index

Feature extraction

```
WITH
  data AS
    (SELECT * FROM [690].[All3col]),
  bounds (minLat,minLong,maxLat,maxLong) AS
    (SELECT min(latitude),min(longitude),max(latitude),max(longitude)
   FROM data),
  binnedSpecies AS
    (SELECT data.species,floor((data.latitude-bounds.minLat)/0.1) AS latBin,
      floor((data.longitude-bounds.minLong)/0.1) AS longBin FROM data, bounds),
  binnedSpeciesCount AS
    (SELECT latBin,longBin,COUNT(species) AS numSpecies FROM binnedSpecies
     GROUP BY latBin,longBin)
SELECT * from binnedSpeciesCount
```

Similarity Indexes

Capture different perspectives of similarity

- **Edit Index**
  - Nb of operations for transforming query q_{k-1} in query q_k.
  - Emphasis in differences
- **Jaccard Index**
  - Emphasis in common parts (relative)
- **Common Fragments Index**
  - Emphasis in common parts (absolute)
- **Common Tables Index**
  - Emphasis in common tables w.r.t. max nb of tables in session
- **Cosine Index**
  - Comparison of vectors in features' space
  - Emphasis in query size

<2,0,0,1,0,0,0,1>
<4,2,0,2,0,0,0,1>

Experiments

- **4 workloads**

Workload	Language	Users	Ground truth	Timestamps	
SQLShare	SQL	Anonymous end-users	✗	✗	
Open	MDX	Master students	✓	✓	[Djedaini et al. 2017]
Enterprise	MDX-like	Developers	✓	✗	[Drushku et al. 2017]
Exam	SQL	Bachelor students	✓	✗	[Kul et al. 2018]

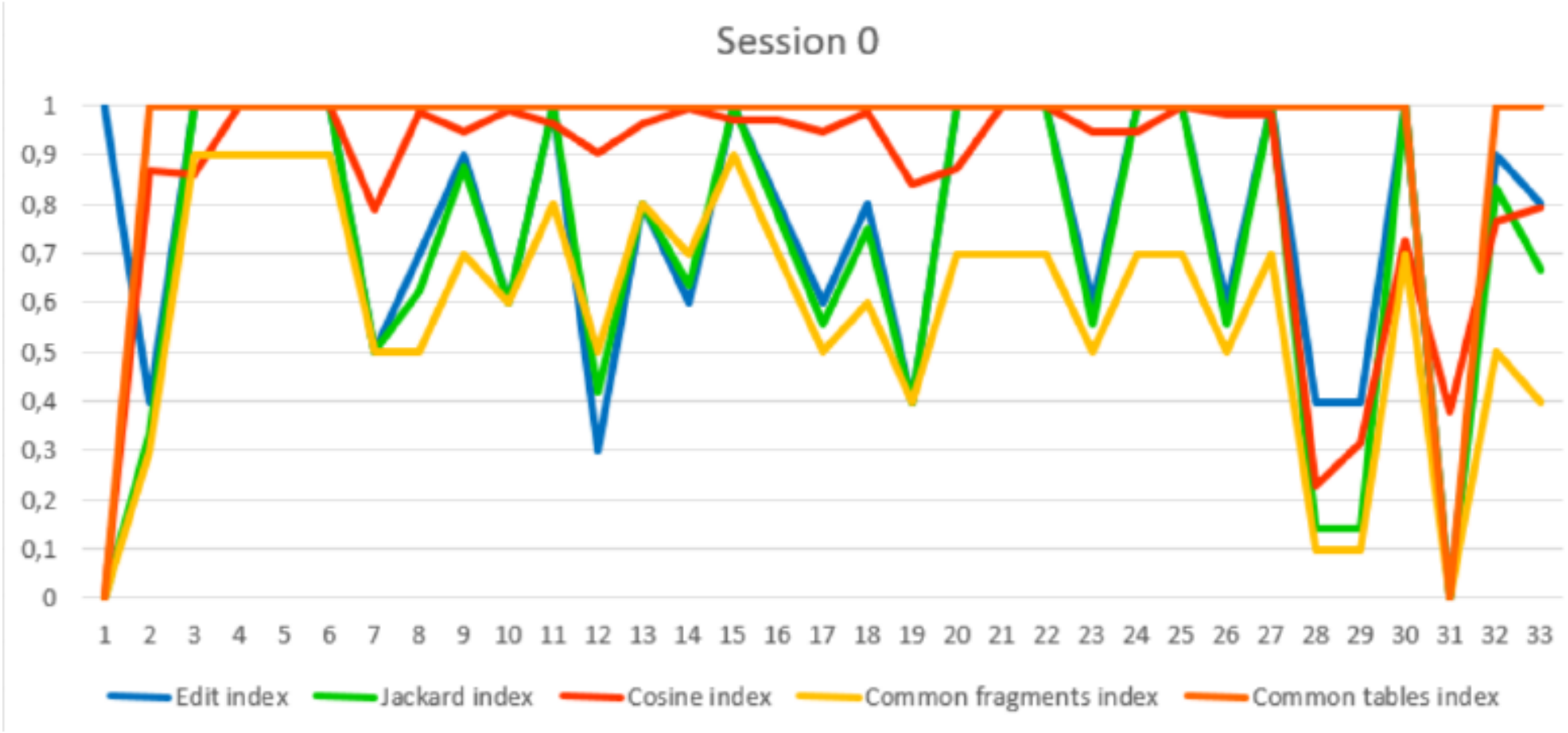
Experiments on SQLShare

Statistics on feature extraction

	Min	1 quartile	2 quartile	3 quartile	Max
Nb projections	1	2	5	10	509
Nb selections	0	0	1	1	83
Nb aggregations	0	0	0	0	49
Nb tables	0	1	1	1	84
Nb common projections	0	0	1	5	509
Nb common selections	0	0	0	1	82
Nb common aggregations	0	0	0	0	48
Nb common tables	0	0	1	1	83
Rel. edit distance	0	2	4	12	1020
Jaccard index	0	0	0.43	0.83	1

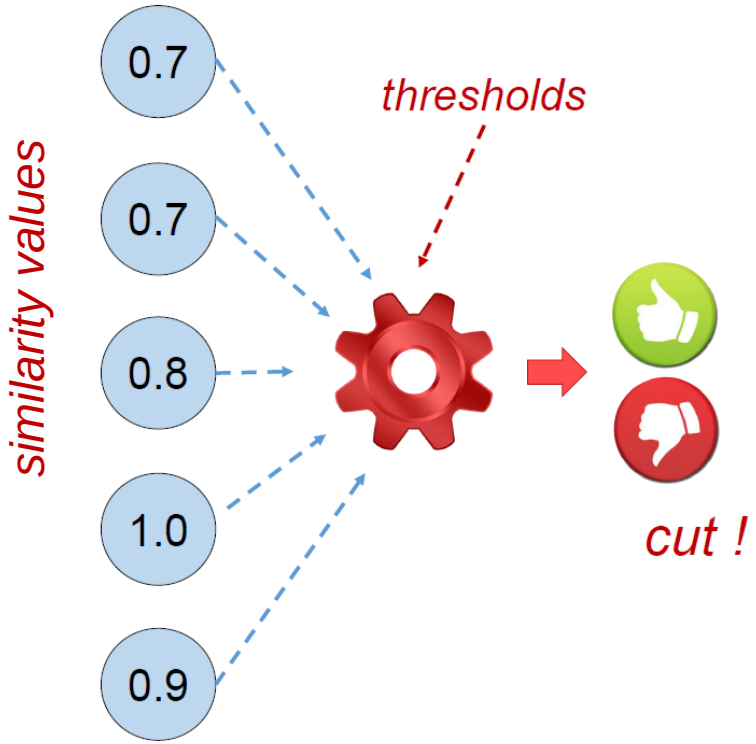
Experiments on SQLShare

Example of similarity values



Experiments on SQLShare

Similarity thresholds



	Edit index	Jackard index	Cosine index	Common fragments index	Common tables index
Min	0,00	0,00	0,05	0,00	0,00
10pc	0,00	0,00	0,68	0,00	0,00
20pc	0,00	0,00	0,72	0,00	0,00
30pc	0,00	0,10	0,81	0,10	0,00
40pc	0,40	0,29	0,89	0,20	0,05
Median	0,60	0,50	0,95	0,30	0,20
60pc	0,80	0,67	0,99	0,50	0,50
70pc	0,80	0,80	1,00	0,60	0,50
80pc	0,90	0,91	1,00	0,90	1,00
90pc	1,00	1,00	1,00	1,00	1,00
Max	1,00	1,00	1,00	1,00	1,00

Experiments on SQLShare

Preliminary segmentation

- Split of the initial 451 sessions in 2 960 explorations
  - Half of sessions were not segmented.
  - Extremely large sessions were very segmented.
- Increase of common fragments. Decrease of edit distance.

	Before segmentation					After segmentation				
	Min	1 quartile	2 quartile	3 quartile	Max	Min	1 quartile	2 quartile	3 quartile	Max
Nb queries	1.00	2.00	6.00	19.75	936.00	1.00	1.00	3.00	6.00	78.00
Avg NCT	0.00	0.66	0.97	1.00	4.00	0.00	0.80	1.00	1.00	83.00
Avg NCF	0.00	2.00	4.20	7.33	306.33	1.00	2.69	5.00	9.50	510.00
Avg RED	0.00	2.30	4.64	8.29	204.73	0.00	1.67	3.00	7.00	267.00
Avg JI	0.00	0.38	0.55	0.71	1.00	0.01	0.43	0.65	0.84	1.00

Experiments with Ground Truth

Characteristics and feature extraction

	Open	Enterprise	Exam	SQLShare
nb sessions	16	24	1	451
nb explorations	28	104	2	
nb queries	941	525	102	10 668
queries/session	58	21	102	24
queries/exploration	34	5	51	
explorations/session	2	4	2	
avg Nb projections	3.62	2.18	1	9.36
avg Nb selections	1.33	0.76	1.57	1.19
avg Nb aggregations	1.34	1.14	0.77	0.41
avg Nb tables	3.28	2.03	3.02	1.50
avg common project	3.16	1.34	0.22	4.90
avg common select.	1.13	0.46	0.07	0.59
avg common aggreg.	1.17	0.77	0.09	0.21
avg common tables	2.97	1.46	2.57	0.85

Experiments with Ground Truth

Evaluation protocol

- **Comparison with ground-truth :**

$< q_0, q_1, q_2, q_3, q_4, q_5, q_6, q_7, q_8, q_9, \ldots >$

Our segmentation → $< 0, 0, 0, 1, 0, 1, 0, 0, 0, \ldots >$

Ground truth → $< 0, 0, 0, 1, 0, 0, 0, 1, 0, \ldots >$

0=cut

1=cut

  - accuracy, precision, recall, f-measure and adjusted rand index (ARI)

- **Threshold setting** : k-percentile in value distribution, k in {0, 5,… 30}
  - Keep threshold that obtains best results

Experiments with Ground Truth

Qualitative results

	Open	**Enterprise**	**Exams**	**Open** (timestamp)
Accuracy	0.98	0.88	0.94	0.97
Precision	1	0.78	0.17	1
Recall	0.42	0.63	1	0.25
F-measure	0.42	0.48	0.17	0.25
ARI	0.75	0.77	0.54	0.75
Best threshold	0	15	5	

The exact break is not always found, while the overall segmentation remains good

Experiments with Ground Truth

- **Correlation between similarity metrics and ground truth**

	Open	Enterprise	Exam
Edit index	0.34	0.62	0.05
Jackard index	0.86	0.73	0.04
Cosine index	0.75	0.32	0.13
Common fragments index	0.86	0.69	0.10
Common tables index	0.90	0.50	0.01

Conclusions

- A proposal for segmenting sequences of SQL queries into meaningful explorations when:
  - only the query text is available (no access to instances)
  - no available timestamps
- based on:
  - a set of simple query features
  - a set of similarity indexes among queries

Both can be extended

- Preliminary results are promising
  - Good results for datasets with ground truth

Perspectives

- Extensions and tunning:
  - Further query features : common fragments w.r.t. near queries, query results
  - Other similarity indexes and thresholds.
- Discard preliminary hypothesis about chronological ordering
  - Test other ways of segmenting, in particular via clustering methods.
- Long term goal : measure the quality of SQL explorations
  - detection of focus/exploratory zones, discovery of latent user intents, recommendation of next queries...
  - for assisting interactive database exploration