Lucerne University of
Applied Sciences and Arts

HOCHSCHULE
LUZERN

Engineering and Architecture
FH Zentralschweiz

Skyline Computation on Commercial Data

Marc Pouly, Michael Galli, Roland Christen, and Stefan Schnürle

SGAICO 2015: Hot Topics in Cognitive Science & Artificial Intelligence
Geneva, Switzerland

Introduction

- Many *different skyline algorithms* exist in the literature.
- Most of their evaluations are based on *synthetic data*.
- In this work, we present a case study of skyline computation on a *representative data set of commercial data* [1].
- Preferences in practice usually include *pairs of differently correlated attributes*, yet almost all experiments in the literature investigate *only pure settings*.

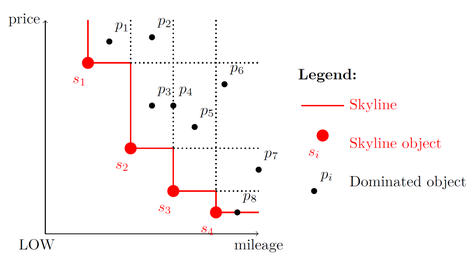
The algorithms inspected are:

- Block Nested Loop (entropy-based window management) [2]
- Divide & Conquer [3]
- Hexagon (lattice-based) [4]
- Scalagon (combination of Hexagon and BNL) [5]

Real Data (i.e., Commercial Data)

- Our data set contains data on 55208 cars [1].
- To each car, 23 attributes are assigned.
  - correlated (e.g., cylinders and engine size).
  - anti-correlated (e.g., mileage and registration date).
  - nearly independent (e.g., mileage and horsepower).
- Outliers countervail correlation effects.
- Cardinalities differ greatly, e.g.:
  - 5988 different values for attribute price.
  - only 17 different values for color.
  - only 6% of all cars are assigned a unique value for price.

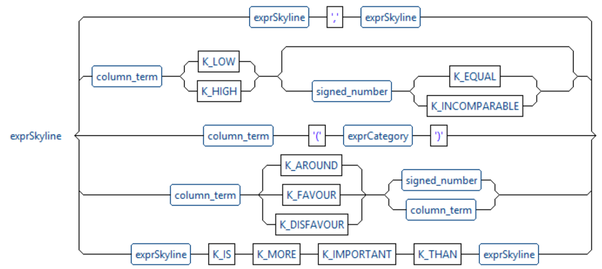
Skylines



- Skyline contains the most *interesting* objects of a data set.
- *Interestingness* according to Pareto dominance paradigm.
- Guarantees *absolute fairness* among all preferences.

prefSQL Framework

```
SELECT          cars.id, cars.title, cars.price, colors.name, bodies.name
FROM            cars
                LEFT OUTER JOIN colors ON cars.color_id = colors.ID
                LEFT OUTER JOIN bodies ON cars.body_id = bodies.ID
SKYLINE OF      cars.price LOW, colors.name ('red' >> OTHERS EQUAL),
                bodies.name ('cabriolet' >> 'limousine' >> OTHERS EQUAL)
```



The preferenceSQL framework implemented in [1] allows for *intuitive specification* of a user's preferences while providing a means to *choose a specific algorithm*.

Experiment Settings and Results

- Set of 17 preferences: numeric (e.g., "low price") and categorical (e.g., color "red » blue » all others").
- 100 random draws from this set, each draw containing a random number of between 3 and 7 preferences.
- 5.13 preferences per run on avg, results see table below.
- *Realistic setting*, since users will most probably choose a mixed set of numeric and categorical preferences.
- *More than 50%* of the runs could not be executed by the Hexagon algorithm (due to large preference cardinalities). Runtime results on Hexagon are therefore incomplete.

	BNL	D&C	Hexagon	Scalagon	Skyline	Min Corr.	Max Corr.
Avg	7	196	2550	175	166	−0.40	0.48
Min	2	45	146	10	1	−0.81	−0.01
Max	96	467	35137	6060	2763	0.01	0.92
Std	11	69	5451	628	331	0.22	0.28

- runtime in milliseconds; avg, min and max runtime of each set of experiments, standard deviation and skyline size.
- Min and max Pearson correlation between any two attributes.

Conclusion

- The results of our measurements *differ significantly* from the results reported on synthetic data in the literature.
- BNL and D&C style algorithms *outperformed lattice-based algorithms* in all our experiments.
- In almost all cases, BNL turned out to be the *best choice* for a practical skyline algorithm.
- The D&C algorithm, that consistently showed only slightly worse runtime compared to BNL, has *great potential for parallelization* on modern microprocessor architectures.
- Hexagon can *hardly be applied* for skyline computation in e-commerce applications.
- We could observe a *strong influence of outliers* in the data set on the performance of skyline algorithms; in contrast to synthetic data, commercial catalogs will almost always contain strong outliers.
- When preference queries are to be computed in concrete commercial applications and on data sets, whose statistical properties have been analyzed, the rich skyline literature with all its investigations on synthetic data *does not provide helpful indications* on which skyline algorithm to apply.

Acknowledgments

We gratefully acknowledge the close and inspiring collaboration with our industry partner Arcmedia AG (www.arcmedia.ch), and we thank our colleagues for valuable remarks and discussions regarding this work.

References

[1] M. Galli, S. Schnürle, R. Arnold, and M. Pouly. (2015) prefsql code repository and experimental setting. [Online]. Available: https://github.com/migaman/prefSQL

[2] J. Chomicki, P. Godfrey, J. Gryz, and D. Liang, "Skyline with presorting," in *ICDE*, U. Dayal, K. Ramamritham, and T. Vijayaraman, Eds., IEEE Computer Society, 2003, pp. 717–719.

[3] S. Börzsöny, D. Kossmann, and K. Stocker, "The skyline operator," in *17th IEEE International Conference on Data Engineering*, 2001, pp. 421–430.

[4] T. Preisinger and W. Kiessling, "The hexagon algorithm for pareto preference queries," 2007.

[5] M. Endres, R. Roocks, and W. Kissling, "Scalagon: An efficient skyline algorithm for all seasons," *DASFAA: 20th Int. Conference of Database Systems for Advanced Applications*, pp. 292–308, 2015.