Reminiscences on Influential Papers

This issue's contributions highlight the impact and educational value of the qualitative and quantitative analysis papers. Enjoy reading!

While I will keep inviting members of the data management community, and neighboring communities, to contribute to this column, I also welcome unsolicited contributions. Please contact me if you are interested.

Pınar Tözün, *editor*
IT University of Copenhagen, Denmark
pito@itu.dk

Vanessa Braganholo
Universidade Federal Fluminense, Brazil
vanessa@ic.uff.br

Juliana Freire, David Koop, Emanuele Santos, and Cláudio T. Silva.
Provenance for Computational Tasks: A Survey.
In Computing in Science and Engineering, Volume 10, Issue 3, pages 11-21, 2008.

While most people choose a primary study for this column, I will be one of the exceptions to the rule and choose a secondary study instead. Secondary studies are especially important when one is learning a new research area. A good one presents and organizes the main concepts and points to the main literature of the area. This is precisely the case of the paper I chose.

When I first started to take an interest in the provenance of scientific experiments, this paper did not exist yet. At that time, I had been recently hired at Universidade Federal do Rio de Janeiro, and I was starting to move out of the ramifications of my thesis subject to something new. Naturally, I was reading lots of papers and finding their relationships on my own, as we all do when we start on a new subject. So, when I found this paper and saw how nicely it put the pieces of the puzzle together, it quickly became one of my favorites.

This is a unique paper in the sense that it does not look like a usual survey. The venue in which it was published works more as a magazine than as a journal, and due to that, the paper gets straight to the point, with few references and a more practical tone. Despite having only 24 formal references, it includes pointers to real systems and practical examples of how they handle provenance. In the same year it was published, it contributed as one of the sources for a tutorial given by Susan Davidson and Juliana Freire at SIGMOD (*Provenance and scientific workflows: challenges and opportunities, SIGMOD 2008*).

Although the field has evolved significantly since the paper's publication, with many new systems and techniques emerging, its core remains valid and highly useful for newcomers. To this day, I recommend this as the first paper my students read when joining my research group.

In addition to its practical applications, this paper has influenced how I guide my students. It has contributed to enforcing my belief that there is always room for good surveys. Since then, I have been encouraging my students to produce good surveys whenever there is none available for the subject they are working on.

Martin Hentschel
IT University of Copenhagen, Denmark
mhent@itu.dk

Roger Weber, Hans-J. Schek, and Stephen Blott.
A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces.
In Proceedings of the 24th VLDB Conference,

pages 194-205, 1998.

In ACM Transactions on Database Systems (TODS), Volume 6, Issue 2, pages 213-226, 1981.

When Pınar asked me to write about a paper that has influenced me, this paper immediately came to mind. It is not that this paper had a major impact on my career, but it is the paper that I have read and thought about most often. I first learned about this work at the VLDB conference in Auckland, New Zealand, in 2008. The paper won the 10-year best paper award. The award session presentation was the most memorable talk of the conference for me.

In the paper, the authors analyze the problems of similarity search in high-dimensional space. They address the "curse of dimensionality," which makes traditional indexing methods inefficient at high dimensions. The authors show that as dimensionality increases, indexing structures like grid-files, quadtrees, R-trees, and clustering methods perform worse than a simple sequential scan. This is because, in high-dimensional space, the data distribution becomes so sparse that efficient partitioning to build indexes is impossible.

What I liked about the paper is that the findings are counterintuitive. Indexing data is bad, scanning data is good, even if you only want to find a single nearest neighbor in the data space. This is the opposite of what I learned in database lectures. Interestingly, the authors show that this applies to all indexing methods, even to methods that may not yet have been invented. They demonstrate this through theoretical analysis, supported by practical results. With this paper, the authors have redefined research on indexing data in high-dimensional spaces.

What I also liked about the paper is that it is easy to read but sometimes hard to understand. While reading, I do not grasp everything. But it is fun to think about the problems and counterintuitive findings. I feel that if I read the paper just one more time, I will understand everything. However, I had to re-read it more than once.

As a result of the findings, scanning data to perform similarity search in high-dimensional space is the only solution. The authors propose a technique to improve scan performance by prefiltering data based on approximations. They call this technique the vector approximation file. In the industry, other techniques to improve scan performance exist, such as pruning data based on min-max statistics, also called block range indexes, zone maps, or small materialized aggregates. For example, Snowflake, my previous employer, makes heavy use of pruning based on min-max statistics to improve query performance. In this sense, the paper has influenced me. I never doubted that scanning is a great solution for addressing problems at scale.