Reminiscences on Influential Papers

Kenneth A. Ross, editor

I continue to invite unsolicited contributions. See `http://www.acm.org/sigmod/record/author.html` for submission guidelines.

Nicolas Bruno, Microsoft Research, `nicolasb@microsoft.com`.

[Improved Histograms for Selectivity Estimation of Range Predicates. Viswanath Poosala, Yannis Ioannidis, Peter Haas, Eugene Shekita. Proceedings of the 1996 ACM SIGMOD international Conference on Management of Data, pages 294–305.]

In early 2000 I was looking for a thesis topic in Databases as a doctoral student at Columbia University. Several papers that I was studying at that point relied on histograms for different purposes. It then seemed a good idea for me to use some time to learn what the state of the art was in the histogram literature. I read some papers on histograms (e.g., equiDepth, end-biased, and serial histograms to name a few) and at first it seemed to me that each one presented a very different approach for cardinality estimation, where the only similarity was the name "histogram".

Then I read the paper by Poosala et al, which studied all aspects of histogram design, the available choices, and how each choice influenced the resulting histograms. This paper was able to characterize all known histograms depending on three design properties (partition class, partition constraint, and sort and source parameters) and two usage properties (intra-bucket approximation of values and frequencies). This taxonomy not only unified different previously proposed techniques, but also allowed to easily identify new design opportunities. Novel histogram techniques, such as MaxDiff and Compressed, were then proposed in paper and evaluated, resulting in better estimates than previous approaches.

After reading this paper, I forgot for a while about the initial topics that relied on histograms and focused on studying histograms themselves. This spawned some publications and finally my thesis dissertation. This paper also showed me that sometimes it is fruitful to step back and try to find a common framework that fits seemingly disparate alternative designs for solving the same problem. Discovering these underlying principles and structures often shed more light on a subject than understanding each alternative separately.

Sam Madden, MIT, `madden@csail.mit.edu`.

[Jim Gray, Paul McJones, Mike Blasgen, Bruce Lindsey, Raymond Lorie, Tom Price, Franco Putzolu, and Irving Trader. The Recovery Manager of the System R Database Manager. ACM Computing Surveys 19(2), pages 223–243, 1981.]

I will be teaching this paper to 50 MIT undergraduates next week — for most of them, it will be their only classroom exposure to core database concepts. Reflections on the significance of this observation for our community or the university aside, this says a great deal about the longevity and accessibility of this paper. The paper clearly and succinctly describes the transactional model and log-based recovery. In doing so, it manages to skirt a fine line: it avoids excessive detail while still conveying many of the subtle nuances that had to be right for the System R recovery manager to provide not only atomicity, but also performance.

One of the paper's most charming aspects is the third section, which consists of a number of reflections on the shortcomings of the original design, written with a candor rarely seen in modern database publications. The authors own up to their realization that System R's hybrid approach to recovery, which combines shadow pages and write-ahead logging, is inferior to pure write-ahead logging. Indeed, because System R relied on this hybrid approach, some readers may view this paper as a simple historical footnote relative to follow-on papers that describe pure write-ahead logging approaches (for example, Aries); those readers would be

mistaken. Both as a generally accessible description of transactions and recovery, and as a snapshot of the tremendous achievement that was System R, this paper deserves to be reread.

Wei Wang, University of North Carolina at Chapel Hill, weiwang@cs.unc.edu.

[R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. SIGMOD 1998, pp. 94–105.]

Clustering has been a classical problem studied in the areas of statistics, pattern recognition, machine learning, and data mining. Before 1998, the curse of dimensionality was addressed mostly from the performance perspective. The mathematic model of clusters always required all dimensions to be included and defined a disjoint partition for a given set of points based on some pre-specified similarity function. This model inevitably suffers from a number of drawbacks including the meaningfulness and interpretability of the discovered clusters in very high dimensional space.

This article first motivated the need for subspace clustering by recognizing that, in high dimensional data, clusters may exist in different subspaces comprised of different combinations of dimensions; and then investigated the problem of finding clusters in subspaces. As the pioneering work on subspace clustering, what impressed me was that it did not only identify a weakness of all previous clustering models but also proposed an efficient method CLIQUE for automatically discovering the appropriate subspaces where better clustering can be achieved. Moreover, the proposed algorithm is rooted in solid information theory. Since its publication, many advances have been made by researchers (including myself) world wide based on the fundamentals it laid down and till now, subspace clustering remains as an active research direction in data mining. Successful applications of subspace clustering can be found in many other disciplines such as E-commerce, information retrieval, and bioinformatics.
