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.

Johannes Gehrke, Cornell University, `johannes@cs.cornell.edu`.

[Leo Breiman, Charles J. Stone, Richard A. Olshen, Jerome H. Friedman. Classification and Regression Trees. CRC Press, 1984.]

Long before data mining appeared on the radar screens of the computer science and statistics research communities, Breiman, Friedman, Olshen and Stone worked on tree-structured approaches to classification and regression problems. Their results are collected in this book which is considered the bible in the field of decision trees. I started reading this book during my second year as a graduate student, and it had a huge influence on my view of the field – and on my understanding of the endeavor of research itself. The book gives not only an in-depth introduction to the CART algorithm, but combines problem motivation with problem solving: The authors lay out a research problem, develop (good and bad) ideas to attack the problems, derive general principles from their ideas, and derive algorithms that implement these ideas.

An example of the depth of the book is the author's treatment of split selection. In the book, split selection is based on the authors' notion of impurity of a set of records, with the goal of split selection to minimize the overall impurity. Instead of jumping right into possible instantiations of such impurity functions and discussing their advantages and disadvantages, the authors first derive general properties that *any* good impurity function needs to have; the instantiation of some impurity functions (such as the gini-index and the twoing criterion) follows naturally from these principles. The lasting impact of this general framework is still seen today, as today's commercial data mining tools all implement (among others) impurity-based split selection methods. The authors guide the reader through similarly inspiring paths when discussing cost-complexity pruning, how to incorporate classification costs, regression tree construction, and decision trees with oblique splits.

This book has had a lasting influence on the field, and Jerry Friedman received the 2002 ACM SIGKDD Innovations Award for his work in this area. The book will stay a classic in decision tree construction, and it is a truly inspiring reading for reader anybody interesting in learning how to do good research.

Jun Rao, IBM Almaden Research Center, `junrao@us.ibm.com`.

[César A. Galindo-Legaria, Arnon Rosenthal. Outerjoin Simplification and Reordering for Query Optimization. TODS 22(1): 43-73(1997)]

First of all, I have to admit there are other papers that have had equal influence on my research. I deliberately picked this one because most others have already appeared in this column before. I first read this paper while doing a summer internship at IBM Almaden Research Center. I was asked to design a practical reordering algorithm for DB2 that can handle outer and anti joins in addition to inner ones. Then I found this TODS paper by Garlindo-Legaria and Rosenthal. It wasn't until I started reading this paper before I realized how dense it was. It not only laid the foundation of outerjoin simplification and association rules, but also covered two flavors of reordering algorithms—one that only considered legal reordering and the other that allowed illegal reordering while providing proper compensation (through generalized outerjoins) to guarantee the correctness. I had to read this paper over and over for three weeks in order to fully appreciate the ideas. Once understanding the paper, I actually found it hard to remove anything from it without losing its value. I was able to complete my assignment successfully that summer, thanks to this paper. Besides serving as a

must-read for people who want to study outerjoins, this paper also taught me how simple algorithms could be developed in a conventional system to support seemingly complicated more advanced features.

Jerome Simeon, Bell Laboratories, simeon@research.bell-labs.com.

[Serge Abiteboul, Sophie Cluet, Tova Milo. Correspondence and Translation for Heterogeneous Data. Proceedings of the 6th International Conference on Database Theory (ICDT'97), Delphi, Greece, January 8-10, 1997. Lecture Notes in Computer Science 1186: 351-363. Springer 1997, ISBN 3-540-62222-5.]

My research career exists because of one paper: "Correspondence and Translation for Heterogeneous Data," which was the result of some work by Serge Abiteboul, Sophie Cluet and Tova Milo during the first half of 1995. It was one of the first papers on semistructured data, and came around the same time as the other two seminal papers in the area ("Programming Constructs for Unstructured Data" by Buneman et al, DBPL'1995, and "Object Exchange Across Heterogeneous Information Sources" by Papakonstantinou et al, ICDE'1995).

I read the first version of the "Correspondence" paper as a fresh PhD. candidate in the first weeks of September 1995, but it was only published two years after that. I do not think it struck people as being such an important paper, but it was the first to give me a complete vision of the technology needed to support Web information exchange and data integration. 8 years after, I am still working on implementing that vision. "Heterogeneous Data" became XML, "Translation" became XQuery, huge work has yet to be done on the related infrastructure. But all of those are details when the vision is given to you...
