Approximate Frequency Counts over Data Streams

Gurmeet Singh Manku
Google Inc., USA
gurmeet@gmail.com

Rajeev Motwani

ABSTRACT

Research in data stream algorithms has blossomed since late 90s. The talk will trace the history of the Approximate Frequency Counts paper, how it was conceptualized and how it influenced data stream research. The talk will also touch upon a recent development: analysis of personal data streams for improving our quality of lives.

1. BIOGRAPHICAL SKETCHES

Gurmeet Manku (1973-) is a software engineer at Google since 2004. He has worked for Infrastructure, Google+ and Ads teams. Presently, he is part of the Google Analytics team which is focused on data mining of clickstream data.

Gurmeet finished his B.Tech in Computer Science at IIT Delhi (1995). He then got his M.S. and Ph.D. from UC Berkeley (1997) and Stanford University (2004) respectively. In between, he worked in the Exploratory Database Group at IBM Almaden Research Center for two years.

Gurmeet has written over 20 research papers in top conferences. His areas of interest have included data stream algorithms, peer to peer systems and data compression.

Rajeev Motwani (1962-2009) was a professor of Computer Science at Stanford University whose research focused on theoretical computer science. He was an early advisor and supporter of companies including Google and PayPal, and a special advisor to Sequoia Capital.

He completed his B.Tech in Computer Science from IIT Kanpur in 1983, got his Ph.D. in Computer Science from U.C. Berkeley in 1988 under the supervision of Richard Karp and joined Stanford soon after U.C. Berkeley. Motwani was one of the co-authors (with Larry Page and Sergey Brin, and Terry Winograd) of an influential early paper on the PageRank algorithm, the basis for Google's search techniques in its early days. He also co-authored another seminal search paper What Can You Do With A Web In Your Pocket with those same authors. He was also an author of two widely-used theoretical computer science textbooks, Randomized Algorithms (Cambridge University Press 1995, with Prabhakar Raghavan) and Introduction to Automata Theory, Languages, and Computation (2nd ed., Addison-Wesley, 2000, with John Hopcroft and Jeffrey Ullman).

Prior to his involvement with Google, Motwani founded the Mining Data at Stanford project (MIDAS), an umbrella organization for several groups looking into new and innovative data management concepts. His research included data privacy, web search, robotics, and computational drug design.

He was an avid angel investor and had funded a number of successful startups to emerge from Stanford. He sat on the boards of Google, Kaboodle, Mimosa Systems, Adchemy, Baynote, Vuclip, NeoPath Networks (acquired by Cisco Systems in 2007), Tapulous and Stanford Student Enterprises among others. He was also active in the Business Association of Stanford Entrepreneurial Students (BASES).

He was a winner of the Gödel Prize in 2001 for his work on the PCP theorem and its applications to hardness of approximation. He served on the editorial boards of SIAM Journal on Computing, Journal of Computer and System Sciences, ACM Transactions on Knowledge Discovery from Data, and IEEE Transactions on Knowledge and Data Engineering.

2. CITATION FROM THE TEN-YEAR-BEST-PAPER AWARD COMMITTEE

This paper [1], one of many on the hot topic of data streams that year (2002), presents algorithms for computing frequency counts exceeding a user-specified threshold. The paper deftly combines theory, algorithms, and experiments, introducing novel algorithms for sticky sampling and lossy counting (though with provably bounded error), which are important for many applications in databases in general, in data mining, in web-server logs, and in networking. A beautifully written paper, it has garnered a truly amazing number of citations over the last decade (a quality shared by some of the other papers appearing in that unusually impactful conference), including a good number just in the last year, a sign that the paper is still quite relevant. One such general concept highlighted in the paper is that of summary data structures with a small memory footprint.

3. REFERENCES

[1] Manku, G. S. and Motwani, R. L. L. Approximate Frequency Counts over Data Streams. *Proc. 28th International Conference on Very Large Data Bases,* 2002, 346-357.

Proceedings of the VLDB Endowment, Vol. 5, No. 12