Very Large Data Bases, VLDB-2016

New Delhi <http://vldb2016.persistent.com/>

VLDB-2016 was an amazing conference that attracted a lot of enthusiastic researchers from both academia and industry. Although database management was the prime focus, there were significant representations from large data analytics, crowdsourcing, graph processing etc.

I have highlighted a few major events below and mentioned a few points-of-interest for our lab at the end.

Keynote 1: Ion Stoica (UCB) spoke about the Spark project from UC Berkeley - its genesis to recently released new version. He started with how recent trends in computing have shifted to iterative processing (such as convergence of optimization algorithms) and motivated the need for in-memory data processing. He further discussed Spark which makes up for the deficiencies in Hadoop (and other previous distributed-computing systems), explained Sparks architecture.

Keynote 2: Anand Rajaraman (RocketShipVC) shared his entrepreneurial experiences with BigData and Apps/softwares to tackle specific problems. He laid emphasis on thinking about newer venues to explore, utilize analytics power and gave the mantra of "Data + X" (substitute for innovative X). Anand also spoke about his current venture - RocketShipVC to connect start-ups with potential venture capitalists efficiently.
I personally liked his vision of creating Cyborgs - personalized intelligent agents by making virtual-personalities and his energetic appeal to be a part of data analytics revolution.

Panel discussion: Will AI eats us all?
The panel discussed why database (DB) community seem to be overlooked because of the recent hype in artificial intelligence (AI). They noted how this image problem was hurting them in the form of lesser funding and lesser graduate students. The panel concluded by agreeing that DB community needs to improve its public perception and join hands with AI community to synergistically build systems of great impact.
This session, as it turned out, was fun filled with lot of sarcastic jokes on ML researchers and I was pleasantly surprised to see database community feeling overshadowed by ML-AI.

Tutorial 1: Machine learning in the real world
Lecture by Rajeev Rastogi (Amazon) on basic concepts of machine learning and hands-on session by Amazon engineers on simple few tasks (they ran through their model codes).

Tutorial 2: Human factors in crowdsourcing
Lecture by Sihem (CNRS) on how important it is to accommodate for social factors in crowdsourcing (because humans aren't machines) to make complete sense of crowd-data. She explained how features such as inter-worker affinity, group size, task diversity, user-experience etc., influence crowdsourced data. There was also discussion on how should the efforts on individuals be aggregated in group based tasks.
I liked how the session made clear distinction between simple micro-tasks vs. collaborative-tasks and handled each case separately.

VLDB 10-year Best Paper Award: The New Casper: Query Processing for Location Services without Compromising Privacy
The authors highlighted the importance of location based services that use dynamic geographical coordinates and the need for customized spatio-temporal databases. For example,

- Restaurants nearby Me - Queries posed by moving requester on stationary locations such as

- Buses near my home - Queries by stationary requester over dynamic locations
- My friends near Me - Queries by moving requester on moving locations

Adding privacy constraints to obtain answers by not revealing the requester's exact location makes the problem harder. The talk made its case for efficiently processing queries over static maps and dynamic GPS coordinates of moving objects under privacy constraints.

Best paper: Compressed Linear Algebra for Large-Scale Machine Learning (IBM Research)

This paper gives techniques for matrices compression and perform linear algebra operations like matrix-vector multiplication to be executed directly on the compressed versions. Their goal was to be able to fit data into main memory for ML algorithms. The performance was comparable to uncompressed version along with significant reduction in memory requirement.

A few papers I found relevant / interesting:

1. PIXIDA: Optimizing Data Parallel Jobs in Wide-Area Data Analytics- Algorithm to minimize bandwidth delay while deploying large distributed data.
2. Cümülön: Matrix-Based Data Analytics in the Cloud with Spot Instances- Smart techniques to bid for Amazon Cloud "spot" instances.
3. Towards Maximum Independent Sets (MIS) on Massive Graphs- Algorithms for MIS such that entire graph cannot be fit into memory. Edges are stored externally.
4. Finding Pareto Optimal Groups: Group-based Skyline- Extends normal definition of single point skyline to group based skylines.
5. SlimShot: In-Database Probabilistic Inference for Knowledge Bases- MLN based inference engine that uses Monte-Carlo based techniques to give formal error guarantees.
6. CLAMShell: Speeding up Crowds for Low-latency Data Labeling- Distributed algorithm to obtain crowd labels with minimum time delay.
7. From Competition to Complementarity: Comparative Influence Diffusion and Maximization- Popularizing two or more products on a social network simultaneously such that they influence each other.

Fun Event: All attendees were taken to Kingdom of Dreams, Gurgoan for a theatrical event, dance and drama. They had great food.

Some useful directions we could explore in our lab:

- **Idea**: We should explore the applicability of "Skyline" concept in ranking entities (based on some importance score). Skyline is n-dimensional surface covering peripheral data-points and are used to answer top-k type of queries.
- **Idea**: Use data-splitting ideas from distributed computing to chuck large data and run ML-optimization on them separately.
- Use compressed algorithms from the Best-paper to perform operations (like Mat-Mat or Mat-Vec multiplication) on memory intensive Knowledge-Graph Tensors.
- For lab's infrastructure, we can refer to papers and implement algorithms for Min-Cut, Max-flow, Page-rank, MLNs that work efficiently on large graphs.
- Most papers on large-data analytics were implemented on Spark and its popularity was evident in VLDB. We can explore Spark based architecture for computation on large datasets.
- If we recognize some system deficiencies (such as repeated data loading) which make our ML-algorithms take longer time on large datasets, we could collaborate with DSL to find customized-system solution.