Notes from IEEE BigDataService 2017

Vishnu S. Pendyala

Abstract - The IEEE Big Data Service 2017 international conference was organized from April 7-9 in San Francisco, USA. The conference attracted researchers from several countries and premier research institues. The editor took quick notes, while attending the conference so that the readership can get a quick overview of the topics covered and those from the first day of the conference are provided below. Full papers can be downloaded from IEEE Xplore soon. Details about the conference itself are at http://big-dataservice.net/.

1 Opening Comments

- Three decades ago, non-conventional methods like Neural Networks and Fuzzy Logic were popular with control systems, but no one had the vision to realize the potential of these technologies.
- Future of Big Data is in applications and services, that's what this conference focuses on.
- Average acceptance rate seems to be around 20%.

2 Keynote #1

Prof. Ling Liu, Georgia Tech, Distributed Data Intensive Systems Lab

IoT and Services Computing: A Marriage made in Big Data

- IoT is a killer app for Big Data.
- Services provide new ways of packaging software - stackable modules.
- Cloud is making everything a service.
- Analytics make the "things" in IoT smart, to make them more responsible.
- Humans cannot right away make out trends from Big Data, but smart devices can.
- Algorithm as a service.
- Tiny computers in everything, including things like the freezer in the refrigerator, to avoid heat shocks to ice creams.
- Big Data, First challenge: Discrete Optimization Problems
- Example of a discrete optimization problem (1st challenge): Where to insert sensors to detect water contamination.
- Most of these problems are NP-hard, so greedy algorithms used as a first attempt of optimization.
- If the utility function is monotonic and submodular, solution theoretically guaranteed 63% accuracy.
- Second Challenge: Deep Learning as a service.
- Convolutional Neural Network is a simplified neural network.
- Every layer learns something and passes what is learnt to the next layer.
- Backward learning is to correct errors.
- http://chronicle.com/article/The-Believers/190147
- Apple's Siri, Google's AlphaGo, Self-driving and face recognition technologies all use the same principles of deep learning.
- In Convolutional Neural Networks, we put everything in a grid, like in map-reduce we create a number of tasks.
- Third and final Challenge: Big Graph Processing - very attractive big data processing algorithms.
- Graph queries and iterative algorithms are two different beasts.
- RDF is a good representation of Natural Language.
- Graph processing is quite demanding on memory resources, so may require news ways of memory allocation, requiring us to rewrite Operating Systems.
- Other way is algorithm based optimization.
- Hama is a popular Apache framework for graph algorithms.
- Smart things make IoT Internet of Services (Services Web).
- IoT and Big Data soon becoming essential utilities like water and electricity, which will be delivered by a services network.

Q&A

- Mobile devices are coming with a number of sensors, but the software using the data from them are still primitive. It is an example of a business model that needs improvement. A lot of opportunity here.
- Dropbox and google drive make their introductory offerings free - that business model is quite successful.
- Today, 15% accuracy is significant in CNN be-

cause humans cannot achieve even 5% accuracy in some cases. That's where RDF and semantic web come into picture. RDF provides an excellent and accurate representation of natural language. The bottleneck in case of RDF is processing. One day when graph processing overcomes its bottleneck, semantic web will be popular.

PAPER 1, 10:30AM

Enhanced Over Sampling Techniques for Handling Imbalanced Big Data Set Classification

- Machine Learning does not work well with imbalanced data sets - where one sample dominates over the other sample.
- How to make Machine Learning work on imbalanced data sets?
- Solution presented is an improvement over SMOTE.
- SMOTE: Synthetic Minority Over-sampling Technique available at: https://www.jair.org/media/953/live-953-2037-jair.pdf

PAPER 2, 11AM

Improved Sentiment Classification by Multimodal Fusion

- Sentiment Analysis (SA) is an aspect of Data Mining.
- Machine Learning techniques are too specific to the problem and are not general enough for the purpose of SA.
- Naive Bayes, SVM and EM represent three different classes of algorithms.
- Data Fusion techniques: Majority Voting, Borda Count (rank based), Ordered Weighted Averaging (OWA), Greatest Max / Min / Product, Maximum Inverse Rank (MIR)
- Validation Technique: K-fold Cross Validation.
- Run the ML algorithms to build binary classifiers and combine the results from the various ML algorithms using the data fusion techniques.
- Uses the lexicon provided by NLTK.
- Uses a 10,000 tweet data set also provided by NLTK.

PAPER 3, 11:30AM

Towards Automatic Linkage of Knowledge Worker's Claims with Associated Evidence from Screenshots

- Use OSX tools like OSXInstrumener to collect data and make the associations.
- Other tools: OpenCV, Google Tesseract, Difflib, BLEU, Jaccard, WordNet.
- Collaborative interaction corpus: http://go.ncsu.edu/cic

KEYNOTE #2

Speaker: Professor Bin Yu, UC Berkeley

Title: Mobile Cloud and Data, One Telekom Perspective

- Prediction Vs Interpretation: Prediction must be interpretable for human retention.
- Lasso is essentially L1 constrained Least Squares.
- Deep Convolutional Neural Networks: http://cs231n.github.io
- Does deep learning resemble the brain function?
- Human brain will always lead the way - humans can do much more than CNNs.
- Deep Dream Patterns show consistency between Lasso and Ridge.
- Superheat plots for visualization of stable deep dream images. R has a package to do this.
- Interpretable models possible through Predictability + Stability + Computability (PSC)
- UCB coming up with a new Data Science major.
- More info: https://drive.google.com/file/d/0B8gpOw0SuKG4cGR1NTZpTzBQRGM/edit and https://drive.google.com/file/d/0B8gpOw0SuKG4NTR5MVJWQjhoc2s/view

PAPER 4, 2:30PM

CaPaR: A Career Path Recommendation Framework

- Mine the resume data and job description and recommend jobs / skills using item-based Collaborative Filtering.

PAPER 5, 2:55PM

IRIS:A Goal-Oriented Big Data Analytics Framework Using Spark for Aligning with Business

- Used Machine Learning techniques running on spark for Business Process Reengineering (BPR).

PAPER 6, 3:30PM

When Rule Engine meets Big Data: Design and Implementation of a Distributed Rule Engine using

- Distributed rule engine using (a) Map-reduce or (b) Message passing for rule matching.
- Rule matching via SparkRE SQL after representing rules as Relational queries.
- Also used Drools and compared the results.
- Datasets from LUBM / OpenRuleBench.

PAPER 7, 4PM

Balanced Parallel Frequent Pattern Mining Over Massive Data Stream

- Three features of data stream: Continuity, unbound, and expiration.

PAPER 8, 4:20PM

Data Allocation of Large-scale Key-Value Store System using Kinetic Drives

- Key-value E.g.: userID (key) userProfile (Value); movieName (key), movie (value)
- Kinetic Drive: World's first ethernet-connected hyper-scale storage, has IP address (instead of SCSI bus address).
- Supports key-value pair using LevelDB and can run key-value operations by itself - easy to scale, plug-n-play.
- Clients use kinetic APIs to work with the drives.

PAPER 9, 4:50PM

Scaling Collaborative Filtering to large–scale Bipartite Rating Graphs using Lenskit and Spark

- Graphs are getting larger and processing is not able to scale.
- Solution is to Partition the graphs.
- What is the best graph partitioning scheme?
- Train a supervised model to predict quality of CF.
- Try several partitioning schemes using structural features of graphs.
- This is the first attempt at using graph partitioning with CF.

PAPER 10, 5:10PM

Small Boxes Big Data: A Deep Learning Approach to Optimize Variable Sized Bin Packing

- Are we making good use of space in the boxes used for packing?
- There are many variations of this problem and all of them are NP-hard.
- Even using just the volume (one-dimension) to optimize the space is NP-hard.
- 3-D optimization time grows exponentially, so we attempt only volume (1D).
- Need to depend on heuristics to bring down the complexity of the algorithm.
- Used 8 heuristics for this solution - none is the best for all situations.
- Customize the heuristics for each individual instance.
- Used Deep Learning to train the model - less feature engineering and auto feature selection.
- Heuristic indicator vector to show how the heuristics performed.