Computational Intelligence and State-of-the-Art Data Analytics

William J. Yeager
Retired
Knowledge Systems Laboratory, Stanford University
byeager@fastmail.fm

Jean-Henry Morin
Institute of Information Service Science
University of Geneva, Switzerland
Jean-Henry.Morin@unige.ch

We are now in the era of, "Do everything on the Internet." The amount of Internet data collected and stored in data warehouses has become explosive. The data sources have increased manifold and in 2020 it is estimated that there are 4.5 billion Internet users, the vast majority of whom are always connected, and 8 billion connected IoT devices in our homes, cities, monitoring critical infrastructures, automobiles, earthquake faults, etc. The "big data" thus generated in 2020 is estimated to be several zettabytes. This data is for the most part, but not exclusively, stored in data warehouses in the Cloud and is website accessible.

Among the long list of practitioners and technologists who must analyze and extract knowledge from this data we find members of large Internet companies like Google, Amazon, Microsoft, and Apple; Telecommunication Companies; governments at all levels; Manufacturing; Advertising and Marking; Heath Care; Biomedical, and Pharmaceutical companies; and their complementary research communities. Each requires the use of a subset of the available state-of-the-art Data Analytic algorithms to extract their desired knowledge. For each such domain specific instance of these algorithms, the knowledge they extract from this data provides the "intelligence" that is indispensable for day-to-day operations, and research.

We must also consider the ever present, dark web where its miscreants use malware to relentlessly attack and profit in billions of dollars annually. These malware attacks can generate gigabytes of data per second on the Internet, as well as Wide-Area and Local-Area-Networks. These data necessitate the use of state-of-the-art Big Data Analytic methods to both detect and analyze attacks in real time data streams, log files, etc. This analysis is necessary to permit practitioners to quarantine malware attacks, as well as build appropriate defenses.

While providing solutions are challenges for Data Analytics professionals, state-of-the-art Data Analytics research in both industry and universities has already produced significant advances to provide intelligent tools which Internet companies are now using to tackle the big data problem. Additionally, new Internet companies are now providing open source，unified, state-of-the-art Data Analytic platforms and technologies, which may be run on Cloud services. These latter unified systems are yielding results from analyzing both real time as well as static, structured and unstructured big data. Salient examples are the open source, Apache Spark, TensorFlow, and DeltaLake projects.

As a consequence, the results of these interdisciplinary combinations of disparate fields of research in both academia and industry have set the stage for the unlimited potential of state-of-the-art Data Analytics to solve some of the world's most difficult problems.

These salient research areas can have contributions from both practitioners of, as well as researchers in both academia and industry in the state-of-the-art Data Analytic fields. These include but are not limited to the following:

1. Artificial Intelligence:

URI: https://hdl.handle.net/10125/71452
978-0-9981331-4-0
(CC BY-NC-ND 4.0)

    a. Machine learning, Knowledge graphs, and Expert systems,
    b. Natural language understanding for unstructured text and voice,
    c. Cyber Security:
        i. Real Time network traffic monitoring and analysis to detect, resolve, and prevent malware attacks,
        ii. Log file analysis to detect, resolve and prevent Malware intrusions.
2. Cloud Services: The power and utility of hosting Data Analytics engines.
3. Data Science: Statistics, data analysis, database theory, and data mining
4. Biomedical research: For example, using intelligence and big data in the search for cures of untreatable diseases, and novel, new uses of existing drugs.
5. Computation: The impact of hardware advances in parallel, distributed computing on Data Analytics software's performance.
6. Operations Research: Algorithms used in Data Analytics.
7. Massive, user based, Internet-wide, distributed computing: SETI@home's BOINC and Astropulse are examples.

Cyber Trust: Distributed trust architectures to insure integrity, authentication, accountability, immutability and transparence.

This year we selected one paper from the candidate papers that we received. The paper is "Deep Learning Through Lens of the Classical SQL" written by Len Du while a student at Australian National University. While data base access with SQL has always been used to provide the data to machine learning algorithms, this paper shows that one can can implement the majority of deep learning algorithms in SQL itself.

Most deep learning frameworks, as well as generic machine learning ones, share a de facto standard of multidimensional array operations, underneath fancier infrastructure such as automatic differentiation. As SQL tables can be regarded as generalizations of multi-dimensional arrays, the author has found a way to express common deep learning operations in SQL, encouraging a different way of thinking and thus potentially novel models. In particular, one of the latest trends in deep learning is the introduction of sparsity in the name of graph convolutional networks, whereas sparsity is taken almost for granted in the database world.

As both databases and machine learning involve transformation of datasets, it's hoped that this work can inspire further such efforts utilizing the large body of existing wisdom, algorithms and technologies in the database field to advance the state of the art in machine learning, rather than merely integrating machine learning into databases.