Research Statement

Djellel Difallah (djellel@nyu.edu)

My work focuses on building systems that embed human intelligence into scalable data processing and analysis methods. In particular, I study the use of paid crowdsourcing with a focus on deadline and quality guarantees. My methods draw inspiration from a variety of areas ranging from distributed systems, expert finding or data management, to psychology and human-resources practices. My goal is to create crowd-powered systems tailored for data analytics tasks, not only by improving the quality of the gathered data but also by developing the next generation of crowd-powered machine learning algorithms and data science tools.

Overview

Human computation is a field that studies the development of methods that combine the scalability of computers in dealing with large volumes of data with the cognitive abilities of the human brain. Such a human-machine synergy is possible thanks to the advent of programmable microtask crowdsourcing platforms, which facilitate the recruitment and compensation of online users. Human computation is used in many areas including machine learning, data management, and information retrieval. For instance, crowdsourcing is often used to collect labels to train supervised machine learning classifiers.

Some of the research questions tackled in human computation are: *How to minimize errors? How to get the results faster?* Or *How to minimize the number of crowdsourced operations?* My Ph.D. work has primarily focused on the effectiveness and efficiency of human computation by investigating two crowdsourcing paradigms (*pull* versus *push* crowdsourcing), and proposing task scheduling and recommendation techniques to obtain the results faster and with higher accuracy. As a postdoc, I focused on the analysis of crowdsourcing platforms to understand their dynamic and how big and involved is the human force behind. This dimension is important to quantify since the promised efficiency of human computation stems from the crowd size and their willingness to participate in a given task.

As a future direction, I plan to concentrate on the area of data analytics where I see tremendous opportunities in harnessing human computation. In this area, the most advanced machine learning algorithms are optimized to find patterns, using sophisticated forms of statistics; however, they still lack the ability to create higher level abstraction models for explaining or justifying their outcome. As humans, we proceed differently: we learn, reason, and make decisions based on model-building. My research will put this capability at the core of novel machine learning algorithms, where crowd workers solve complex problems (beyond microtasks) independently and collaboratively.

Past Research

In my thesis work, I investigated methods to combine human computation and scalable data processing systems. Furthermore, I built and evaluated real systems to improve efficiency and effectiveness metrics in human-machine based computation. One of the main ideas supported by my research is to depart from a crowdsourcing platform as a marketplace (pull-crowdsourcing) and adopt a push-crowdsourcing methodology where the tasks are assigned to available workers automatically upon request. This execution pattern not only allows the system to have a fine-grained control over who performs what, it further presents new possibilities to load balance the available workforce across multiple requesters who have different budgets, workloads, and expectations.

In the first work, my co-authors and I used probabilistic inference models that combine crowd responses with the output of an automatic classifier. This work was applied to the tasks of entity extraction from text and instance matching across knowledge graphs [1, 2]. One of the opportunities that I identified and pursued in subsequent work was to selectively route specific tasks to workers who are likely to provide better answers given their personal knowledge. For example, a basketball fan is more likely to recognize NBA players in an image labeling task. Instead of letting the system randomly dispatch the tasks, I developed a task recommender system for workers based on their social media profiles. I showed for instance that there is a correlation between the quality of the responses and the social interests of the workers [3].

From an efficiency point of view, I explored monetary bonuses as an incentive for retaining workers longer on a set of tasks to improve the execution speed. I showed that this model presents two advantages: It minimizes the stall times incurred when waiting for new workers, and workers tend to produce better results as they become more familiar with the task [4].

The final component of my research was the study of scheduling algorithms for tasks execution on a crowdsourcing platform. In fact, a platform like Amazon Mechanical Turk (AMT) has thousands of requesters running hundreds of thousands of tasks with tens of thousands of workers continuously [5]. To improve the utility of such a system, I proposed the use of scheduling algorithms that handle the distribution of the tasks to the workers automatically. The achieved goal was to reduce task starvation and thus improve fairness across requesters [6].

During my PhD years, I have collaborated with over 20 researchers from EPFL, NYU and CMU, and industrial partners such as Microsoft, IBM Research, Paradigm4, and Verisign. My research resulted in 17 peer-reviewed papers published at prominent conferences and in journals such as WWW, VLDB, HCOMP, and SIGMOD. I was honored to receive the Best Thesis Award in Computer Science at the University of Fribourg in 2015, for my PhD thesis entitled "Quality of Service in Crowd-Powered Systems."

As a postdoctoral researcher at the eXascale Infolab, I contributed to two successful grant proposals submitted to the European Commission to fund an H2020-ICT project (Fashion-Brain project[1]) and Swiss Commission for Technology and Innovation (CTI).

Current Research

I have recently taken a Data Science Fellow position at the New York University – Center for Data Science, where I collaborate with data scientists of interdisciplinary backgrounds and with machine learning researchers. My first collaboration resulted in a research (currently under review) that presents an analysis of the population dynamics and demographics of Amazon Mechanical Turk workers. We used techniques from the field of ecology to understand the size and dynamics of the underlying population. We also demonstrate how to model and account for the inherent selection biases in a crowdsourced study. In a similar work, we focused on a different population of workers who perform edits on Wikidata, a crowdsourced open Knowledge Graph. Here, the focus is on the analysis of worker life time analysis, and activity prediction based on activity logs collected from the past five years.

Other Projects

- **A Data-Driven Analysis of Amazon Mechanical Turk:** My co-authors and I collected and processed publicly visible data from AMT over a five years period. Our goal was to investigate the workload evolution on AMT over the years. The result was a long-term data-driven market analysis of AMT where we computed key growth statistics and identified the features that dictate the tasks' completion time [5]. As future work, we plan to quantify the supply and demand phenomena in the marketplace using further collected fine-grained data.

- **Reservation-Based Scheduling:** During my internship at Microsoft's Cloud Information Services Lab, I developed new scheduling algorithms for large clusters of machines running millions of individual tasks per day [7]. My work there inspired my subsequent research endeavors; I adapted scheduling algorithm used in enterprise cloud computing to crowdsourcing platforms.

- **OLTP-Bench a Testbed for Benchmarking:** Together with a group of other PhDs, postdocs, and professors from several universities, we built an opensource testbed for benchmarking relational databases [8]. The framework is designed to produce variable rate and variable mixture load against any JDBC-enabled relational database. OLTP-Bench is now used by many companies and research institutions, including Oracle, Facebook, Microsoft, MIT, UC Berkeley, and Cornell. One of my short-term goals is to develop a similar framework for benchmarking crowd-powered systems through real-world data simulation.

Future Research: Crowd-Powered Data Analytics

The direction of my future research will be aimed at developing crowd-powered systems for data analysis tasks. The key idea is to embed the crowd in data science endeavors, not as a mere data collection or labeling alternative, but at the core of the algorithms and the data exploration process. I expect human computation to improve machine learning not only by creating better predictive models but also by explaining and justifying their output (introspection).

I propose to pursue a research agenda centered around building a crowd-powered data analytics stack. This will include machine learning and data mining algorithms that can leverage human computation, in addition to interactive tools that crowd workers can use either independently and/or collaboratively. In detail, I will be exploring the following themes:

[1] https://fashionbrain-project.eu/

- **Data Mining with the Crowd:** Identifying interesting patterns in large amounts of data is complex both for humans and machines alike. Both approach the problem quite differently. Humans proceed top-down, by first formulating likely scenarios and then trying to find supportive information from the data (see [9] for such a crowd-based approach). In contrast, data mining algorithms, like Apriori [10], proceed bottom-up, exhaustively listing all possibilities and constructing longer rules, while discarding unsupported cases after each iteration. I intend on exploring crowd-powered data mining algorithms starting from the case of association rule mining. Unlike Apriori, which relies exclusively on heuristics, I will selectively question the crowd to keep promising sub-rules. The challenges in that context are (a) *How to minimize the number of inquiries?* And (b) *What are the use cases for such an approach?*

- **Crowdsourcing the Feature Engineering Process:** Data scientists with domain expertise can create successful predictive models by identifying the features and the data that can be used for training. However, when they lack the domain knowledge, the data scientists proceed by inspecting data samples, making predictions, and improving their models iteratively. I will investigate how a similar iterative approach can be carried out by crowd workers. This direction shows promise because crowdsourcing has recently been used to suggest (and label) predictive features for binary classification tasks for images and text [11]. I intend to use crowdsourced workflows composed of feature suggestion, labeling, and testing tasks to (a) tackle a broader set of classification problems and (b) speed-up the feature engineering process. The challenge will be to develop workflow optimization techniques that reduce the cost of this process.

- **Crowdsourced Active Learning:** Active learning is a semi-supervised machine learning technique that, given a budget, selectively picks a data item that an oracle is asked to label. While crowdsourcing can be used to obtain the labels faster, I propose to improve the item selection itself by using the crowd. Here, I would like to investigate batching techniques where the crowd worker receives multiple items and decides which one is more interesting to label. This can reduce the labeling cost and/or produce better models. Additionally, making the decision about which label to gather might depend on which worker is available, and hence, I intend to use probabilistic inference models to select items accordingly.

- **Collaborative Macro-task Crowdsourcing for Data Analytics:** The previously discussed research ideas depart from the classic form of microtask crowdsourcing and promote the idea of giving more context about the problem to the crowd, and, therefore, potentially granting access to more data. For example, a worker might request more samples to judge a feature (or an association rule) interesting. The required samples can be random or follow certain criteria expressed by the worker. Macro-tasks raise multiple research questions including (a) *How to price a macro-task?*, (b) *How to design interfaces for collaborative data analysis?* And (c) *How to prevent information leakage?* Additionally, the functional aspects such as the development of simple data visualization and exploration tools for the crowd need to be investigated.

My future research directions aim at pushing the boundaries of human computation by exploring novel techniques that involve the crowd in solving complex problems, as individuals or as collaborative groups. Hence, I expect the outcome to have a broad impact on a number of areas where human guided data analysis can make a significant difference. To achieve this goal, input from multiple fields of computer science and related disciplines will be critical. I look forward to the opportunity to collaborate with researchers in machine learning and visualization, in addition to researchers in psychology, cognitive and social sciences.

References

[1] Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. Zencrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proceedings of the 21st World Wide Web Conference 2012, WWW 2012, Lyon, France, April 16-20, 2012*, pages 469–478, 2012.

[2] Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. Large-scale linked data integration using probabilistic reasoning and crowdsourcing. *VLDB J.*, 22(5):665–687, 2013.

[3] Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. Pick-a-crowd: tell me what you like, and i'll tell you what to do. In *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013*, pages 367–374, 2013.

[4] Djellel Eddine Difallah, Michele Catasta, Gianluca Demartini, and Philippe Cudré-Mauroux. Scaling-up the crowd: Micro-task pricing schemes for worker retention and latency improvement. In *Proceedings of the Seconf AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2014, November 2-4, 2014, Pittsburgh, Pennsylvania, USA*, 2014.

[5] Djellel Eddine Difallah, Michele Catasta, Gianluca Demartini, Panagiotis G. Ipeirotis, and Philippe Cudré-Mauroux. The dynamics of micro-task crowdsourcing: The case of amazon mturk. In *Proceedings of the 24th International Conference on World Wide Web, WWW 2015, Florence, Italy, May 18-22, 2015*, pages 238–247, 2015.

[6] Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. Scheduling human intelligence tasks in multi-tenant crowd-powered systems. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, pages 855–865, 2016.

[7] Carlo Curino, Djellel Eddine Difallah, Chris Douglas, Subru Krishnan, Raghu Ramakrishnan, and Sriram Rao. Reservation-based scheduling: If you're late don't blame us! In *Proceedings of the ACM Symposium on Cloud Computing, Seattle, WA, USA, November 03 - 05, 2014*, pages 2:1–2:14, 2014.

[8] Djellel Eddine Difallah, Andrew Pavlo, Carlo Curino, and Philippe Cudré-Mauroux. Oltp-bench: An extensible testbed for benchmarking relational databases. *PVLDB*, 7(4):277–288, 2013.

[9] Yael Amsterdamer, Yael Grossman, Tova Milo, and Pierre Senellart. Crowdminer: Mining association rules from the crowd. *PVLDB*, 6(12):1250–1253, 2013.

[10] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *VLDB'94, Proceedings of 20th International Conference on Very Large Data Bases, September 12-15, 1994, Santiago de Chile, Chile*, pages 487–499, 1994.

[11] Justin Cheng and Michael S. Bernstein. Flock: Hybrid crowd-machine learning classifiers. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW 2015, Vancouver, BC, Canada, March 14 - 18, 2015*, pages 600–611, 2015.