Improving Data Management Applications Using Microtask Platforms

Jiannan Wang
Computer Science, Tsinghua University
wjn08@mails.tsinghua.edu.cn

Reynold S. Xin
AMPLab, UC Berkeley
rxin@cs.berkeley.edu

ABSTRACT

Many data management problems are inherently vague and hard for algorithms to process. Take for example entity resolution, also known as record linkage, the process to resolve records for the same entity from heterogeneous sources. Properly resolving such records require not only the syntactic structure of the data, but also contextual semantics that are hard for machines to understand. To properly perform such data management tasks requires human inputs for providing information that is missing from the structured data that machines can read, for performing computationally difficult functions, and for matching, ranking, or aggregating results based on fuzzy criteria.

The rise of microtask crowdsourcing platforms, e.g. Amazon's Mechanical Turk, provides a unique opportunity to integrate human inputs into the algorithmic data flow. There are two recent work, CrowdDB [1] and CrowdER [2], that combine human inputs with machine computation to answer otherwise unanswerable queries.

- **CrowdDB** extends SQL's Data Definition Language to allow the crowd as an input source to provide data that are absent in the database. It can be used to answer questions such as which picture is the best looking.
- **CrowdER** introduces a hybrid human-machine workflow that uses algorithmic approaches to weed out the obviously non-matching pairs, and only sending the remaining pairs to humans for further high-quality confirmation.. It achieves higher quality in entity resolution than state-of-the-art algorithmic approaches.

CrowdDB and CrowdER are only two examples of leveraging microtask platforms to improve traditional data processing flow. We expect this to be a fruitful research area for the many years to come and enable more powerful data management applications.

BODY

Judiciously leveraging microtask platforms can enable more powerful data management applications that are otherwise impossible.

REFERENCES

[1] M. J. Franklin, D. Kossmann, T. Kraska, S. Ramesh, and R. Xin. Crowddb: answering queries with crowdsourcing. In *SIGMOD Conference*, pages 61–72, 2011.

[2] J. Wang, T. Kraska, M. J. Franklin, and J. Feng. Crowder: Crowdsoucing entity resolution. *PVLDB*, 5(10), 2011.

Volume 1 of Tiny Transactions on Computer Science