Crowdsourcing Applications and Platforms: A Data Management Perspective

AnHai Doan[1,2], Michael J. Franklin[3], Donald Kossmann[4], Tim Kraska[3]

[1]University of Wisconsin, Madison [2]@WalmartLabs [3]University of California, Berkeley [4]ETH Zurich

1. INTRODUCTION

Over the past decade, crowdsourcing has emerged as a major problem-solving and data-gathering paradigm on the World-Wide Web. Well-known examples of crowdsourcing include Wikipedia, Linux, Yahoo! Answers, YouTube, Mechanical Turk-based applications, and much effort is being directed toward developing many more.

As is typical for an emerging area, this effort has appeared under many names, including peer production, user-powered systems, user-generated content, collaborative systems, community systems, social systems, social search, social media, collective intelligence, wikinomics, crowd wisdom, smart mobs, mass collaboration, and human computation. The topic has also been discussed extensively in books, popular press, and academia (e.g., [17, 18, 16, 19, 13, 3, 5, 8, 15, 11, 19, 10, 7]).

This extensive attention, as well as the many successes of crowdsourcing, has generated much recent interest in the database community. Recent projects on crowdsourcing in the database community include the CrowdDB project at Berkeley and ETH Zurich [8], the Scoop project at Stanford and the University of California, Santa Cruz [15, 14], the crowdsourcing project at MIT [11], the Cimple/DBLife project at Wisconsin [5] (and the earlier MOBS project at Illinois [12]), the PSOX project at Yahoo! Research [2], the crowdsourcing project at NYU [1], the Computational Journalism project at Duke [6], the dataspace and pay-as-you-go data integration project at Google, Berkeley, and Portland State University [10], the Fusion Tables project at Google [9], the Intelligence in Wikipedia project at Washington [19], and many others (e.g., [4] as well as many efforts in crowdsourcing the construction and maintenance of the Semantic Web).

Recent panels on crowdsourcing in the database community include "Web 2.0 and Databases" at VLDB-07 and "Crowds, Clouds, and Algorithms: Exploring the Human Side of Big Data Applications" at SIGMOD-09. At the same time, many companies have successfully used crowdsourcing to solve a broad range of data management problems.

Proceedings of the VLDB Endowment, Vol. 4, No. 12

The above examples, while by no means exhaustive, clearly demonstrate the emerging strong interest and momentum of crowdsourcing in the database community. From a data management perspective, crowdsourcing allows, for the first time, large-scale and on-demand invocation of human input for data gathering and analysis. Data gathering can be done implicitly, through crowdsourced sensing and on-line behavior, or explicitly, by sending targeted information requests to the crowd. From an analytics perspective, human input can be used to address computationally difficult tasks such as entity resolution, schema matching, object recognition, outlier detection, subjective comparisons (such as fuzzy matching, classification and ranking), and contextual analytics bringing multiple data sources and perspectives.

Of course, crowdsourcing also brings new challenges to data management, including quality assessment and improvement, latency, scheduling, cost optimization, privacy, and social issues. These issues and others lead to a wealth of research topics around systems, semantics, and user interfaces. The goals of this tutorial then are

- to spark further interest and contribute to the gathering momentum of crowdsourcing in our community,
- to introduce database researchers to established and cutting-edge techniques in crowdsourcing,
- to foster more discussions on applying crowdsourcing to data management problems, and
- to identify specific data management related research challenges in crowdsourcing.

2. TUTORIAL OUTLINE

The tutorial will cover the following topics.

2.1 Motivations, Definitions, and Evolution

We begin by describing various notions of crowdsourcing. We distinguish it from other crowd-centric notions such as crowd management. Next, we outline the evolution of crowdsourcing on the Web. Finally, we discuss why crowdsourcing has received a recent surge of attention, both outside and in the database community.

2.2 Crowdsourcing Applications

Many crowdsourcing applications have been developed on the World-Wide Web. Prime examples include book review sites, YouTube, Flickr, ehow.com, Quora, Swivel, Wikipedia,

Demand Media, ESP, and Mechanical Turk-based applications. We discuss how to unify these disparate applications, and the challenges that they share.

- We identify important dimensions along which we can classify crowdsourcing applications.
- We present a classification of basic crowdsourcing applications on the World-Wide Web.
- We discuss how crowdsourcing applications on the Web have combined the above basic types.
- We discuss the fundamental challenges in building these applications: How to solicit users? What can users contribute? How to combine user contributions? And how to evaluate users and contributions? We discuss established and cutting-edge solutions to these challenges.

2.3 Crowdsourcing Platforms

Crowdsourcing application (as discussed above) are built using private or public platforms. Over the past decade, many such platforms have been developed by many companies. Examples include Mechanical Turk, Turkit, Mob4hire, uTest, Freelancer, eLance, oDesk, Guru, Topcoder, Trada, 99design, Innocentive, CloudCrowd, and CloudFlower. In this part of the tutorial we discuss and classify these platforms.

- We begin by discussing Mechanical Turk, a concrete and well-known crowdsourcing platform, to give the audience a feel for how such a platform operates.
- We then discuss a range of other major crowdsourcing platforms, and compare them to Mechanical Turk. We discuss a set of dimensions, then use them to classify crowdsourcing platforms.
- We discuss fundamental challenges that arise when designing and operating crowdsourcing platforms, and efforts in both academia and industry to address these challenges.

2.4 Crowdsourcing for Data Management

So far we have discussed the general landscape of crowdsourcing applications and platforms. We now zoom in and discuss how crowdsourcing has been applied to a range of data management applications, and the challenges and opportunities in doing so.

- We quickly survey a range of data management applications that have benefited from crowdsourcing. Prime examples include schema matching, building structured databases (over unstructured data), data integration, answering SQL queries, graph search, and understanding social media.
- We discuss the challenges that arise in the above applications. Prime examples include how to solicit users, what they can contribute, how to combine their contributions, how to manage quality, open versus close worlds, query semantics, query execution, optimization, and user interfaces.

2.5 How to Start in this Area

We wrap up the tutorial by discussing how researchers new to crowdsourcing can start in this area. We discuss, among others, how to evaluate proposed solutions, using both real and synthetic users. And we highlight further resources on crowdsourcing beyond the current tutorial.

3. REFERENCES

[1] Crowdsourcing project at NYU. `http://pages.stern.nyu.edu/~panos`.

[2] The PSOX project. `http://research.yahoo.com/node/498`.

[3] Wikipedia and artificial intelligence: An evolving synergy, 2008. AAAI-08 Workshop.

[4] K. Belhajjame, N. Paton, A. Fernandes, C. Hedeler, and S. Embury. User feedback as a first class citizen in information integration systems. In *CIDR*, pages 175–183, 2011.

[5] X. Chai, B. Vuong, A. Doan, and J. Naughton. Efficiently incorporating user feedback into information extraction and integration programs. In *SIGMOD*, pages 87–100, 2009.

[6] S. Cohen, C. Li, J. Yang, and C. Yu. Computational journalism: A call to arms to database researchers. In *CIDR*, pages 148–151, 2011.

[7] A. Doan, R. Ramakrishnan, and A. Halevy. Crowdsourcing systems on the World-Wide Web. *Communications of the ACM*, 54(4):86–96, 2011.

[8] M. Franklin, D. Kossman, T. Kraska, S. Ramesh, and R. Xin. CrowdDB: Answering queries with crowdsourcing. In *SIGMOD*, pages 61–72, 2011.

[9] H. Gonzalez, A. Halevy, C. Jensen, A. Langen, J. Madhavan, R. Shapley, W. Shen, and J. Goldberg-Kidon. Google fusion tables: web-centered data management and collaboration. In *SIGMOD*, pages 1061–1066, 2010.

[10] A. Halevy, M. Franklin, and D. Maier. Principles of dataspace systems. In *PODS*, pages 1–9, 2006.

[11] A. Marcus, E. Wu, and S. Madden. Crowdsourcing databases: Query processing with people. In *CIDR*, pages 211–214, 2011.

[12] R. McCann, W. Shen, and A. Doan. Matching schemas in online communities: A web 2.0 approach. In *ICDE*, pages 110–119, 2008.

[13] R. Mihalcea and T. Chklovski. Building sense tagged corpora with volunteer contributions over the web. In *RANLP*, pages 357–366, 2003.

[14] A. Parameswaran and N. Polyzotis. Answering queries using humans, algorithms, and databases. In *CIDR*, pages 160–166, 2011.

[15] A. Parameswaran, A. Sarma, H. Garcia-Molina, N. Polyzotis, and J. Widom. Human-assisted graph search: It's okay to ask questions. In *VLDB*, 2011.

[16] H. Rheingold. *Smart Mobs*. Perseus Publishing, 2003.

[17] J. Surowiecki. *The Wisdom of Crowds*. Anchor Books, 2005.

[18] D. Tapscott and A. D. Williams. *Wikinomics*. Portfolio, 2006.

[19] D. S. Weld, F. Wu, E. Adar, S. Amershi, J. Fogarty, R. Hoffmann, K. Patel, and M. Skinner. Intelligence in wikipedia. In *AAAI*, pages 1609–1614, 2008.