Cite as:
Lukyanenko, R. and Parsons J. (2015). Beyond Task-Based Crowdsourcing Database Research. *AAAI Conference on Human Computation & Crowdsourcing (AAAI HCOMP)*, San Diego, CA.

Beyond Task-Based Crowdsourcing Database Research

Roman Lukyanenko
Florida International University, Miami, FL
rlukyane@fiu.edu

Jeffrey Parsons
Memorial University of Newfoundland, St. John's, NL
jeffreyp@mun.ca

Abstract

The emergence of crowdsourcing as an important mode of information production challenges established approaches to data management. Crowdsourcing has attracted increasing research attention; currently, however, the database research community primarily focuses on crowdsourcing research that can be termed task-based. In this paper, we suggest to broaden this effort to include another important type of crowdsourcing, which we term surveillance-focused. We consider the challenges in this domain, review approaches to data modeling for crowdsourcing, and suggest directions for future research.

Introduction

Driven by a sustained interest in using the availability, skill, and interest of ordinary people, the data management community has increasingly adopted crowdsourcing as an exciting research topic. However, crowdsourcing studies remain somewhat narrowly focused - emphasizing using the crowd to perform small, well-defined tasks. We argue there is a largely untapped research potential in harnessing crowds in broader and less precisely prescribed activities.

A major trend in current research is to investigate uses for crowds as links in a larger **technological chain** where people are seen primarily as problem-solvers in small, independent and well-defined tasks. In this context, Amsterdamer et al. (2015) likens crowds to "an external (and very slow, potentially unreliable) hard drive" that can be queried on demand. We term this stream ***task-based crowdsourcing***.

In task-based crowdsourcing, researchers are attracted to environments such as Amazon Mechanical Turk or CrowdFlower, in which designers can pose typically small ("micro", "tiny") tasks to a crowd of workers often for a payment. These tasks can be sent to innovative database engines that combine traditional SQL statements with user-defined functions (Franklin et al. 2011; Park et al. 2013). Typical small tasks include classification of items, providing missing values, sorting and filtering records, comparing items; they also may include answering open-ended questions (e.g., Amsterdamer et al. 2015; Franklin et al. 2011; Lasecki et al. 2015; Park et al. 2013). Small tasks are typically stripped of broader organizational context and underlying objectives, and can be treated as stand-alone autonomous problems. While special skills may be required to complete a task, broader understanding of the projects or sponsoring organizations is not expected.

While research on task-based crowdsourcing remains important, database knowledge can be expanded through deeper understanding of another major, but less studied model of crowdsourcing, which we term ***surveillance-focused***, wherein *organizations harness human perceptive and information-gathering abilities to make sense of the environment in which organizations operate*. Examples of this type of crowdsourcing include community mapping, crisis management, civic engagement, corporate market surveillance or online citizen science. These are typically full-fledged feature-rich projects allowing people to report on natural or social phenomena they experience in the course of daily life, as well as interact with each other and the organizational sponsors. For example, Cornell University launched eBird (www.ebird.com) to collect bird sightings from bird-watchers across the globe to generate data for their ornithology research program. Data collection in this project is well-structured and involves populating pre-specified fields (e.g., selecting a biological species observed).

Unlike task-based crowdsourcing that places crowds amidst a technological chain to be intelligent mediators to machine tasks**,** surveillance crowdsourcing conceptualizes ordinary people as integral elements of the organizational **information chain** that links the organization with external environment. In this setting, database researchers are confronted with the need to produce solutions that are both technologically sound, as well as effective at unlocking the potential of people as sensors of their surroundings. A key aspect of harnessing the crowd in this manner is recognizing that tasks may be broad and loosely-defined. In this setting, determining data requirements and designing data collection systems accordingly is a critical challenge.

A major feature of surveillance-focused crowdsourcing is open participation. Typical users or contributors are ordinary people, often lacking subject matter expertise and possessing diverse domain views. Whereas many task-based crowdsourcing platforms may constrain who gets to do a task (e.g., by eliminating members with low reputation, requiring a certain skill, or pruning data post-hoc based on known user attributes), this is often impossible when participation is democratic and anonymous. As sponsors are unable to fully determine the domain expertise of crowd providers, it is difficult to design appropriate structures (e.g., relations) that are congruent with the views of all users. For example, consider a project that collects information about wildlife such as an eBird.org, iSpotNature.org, or Treezilla.org. Approaching data modeling in these projects from the traditional perspective would produce a unified set of relations (tables) to structure information collection. It is common to base such relations on the data needs of the sponsoring organization (here, academics looking to use the data). In natural history this often results in data collection at the biological species level (e.g., Lukyanenko, Parsons, and Wiersma 2014; Crall et al. 2010). Yet focusing on this level of granularity may marginalize, bias, or exclude valuable conceptualizations of those users lacking sufficient expertise and familiarity with this classification level.

In many projects, the phenomena about which users supply data may be available *only* to the original contributor. For example, the objects of interest (e.g., birds, animals, cosmic events) may be fleeting with a very short exposure time. In such cases, it is extremely difficult to exploit redundancy in the crowds (e.g., Franklin et al. 2011; Sheng, Provost, and Ipeirotis 2008), and the challenge is to get the most out of a single data point. The anonymous nature of many projects further precludes seeking clarification or additional information unless such a query can be formulated before the contributor leaves the project.

Surveillance projects often have loosely-defined goals to draw "unanticipated benefits" from crowd ingenuity. This suggests that databases need to model multiple, evolving and unexpected uses or even be use-agnostic.

As no established data model for surveillance-focused crowdsourcing exists, major crowdsourcing projects, such as eBird, CitySourced, EpiCollect, are based on traditional (e.g., relational) implementations. In contrast, some researchers argue the relational model has a negative impact on information quality and user participation and may be generally inappropriate for this domain (Lukyanenko, Parsons, and Wiersma 2014). One possible solution is storing user input in an unstructured form (e.g., as free-form text). However, without systematic data production in the direction required by project sponsors, the yield of useful information from free-form data collection is likely to be dismal. A hybrid modeling solution (e.g., based on flexible data models combined with unstructured text) is potentially better suited here.

Another important issue is spatial and temporal data sparsity. For example, if an organization (e.g., city of Miami) is to manage city-wide services based on a crowdsourcing data set (e.g., citysourced.com), it needs to somehow "impute" gaps necessarily arising when data is not collected systematically. A question arises whether it is possible to anticipate such gaps and correct them in real-time before contributor leaves. So far research has demonstrated the significant extent of this problem but offers few solutions (and none to our knowledge dealing with databases).

Crowdsourcing opens many opportunities to develop innovative solutions to support knowledge acquisition from diverse and heterogeneous crowds. The work in the microtask markets has already resulted in promising technologies that improve schema design, query execution and optimization, and support innovative database and interface integration. We call on the database community to increase attention to another type of crowdsourcing, as exemplified by (but not restricted to) surveillance activities performed by amateurs on behalf of an organization. While many challenges of task-based crowdsourcing apply to this context also, the holistic nature of surveillance-focused crowdsourcing projects create challenges that call for ***novel integrated data management solutions tailored to this domain***.

References

Amsterdamer, Y., S.Davidson, A. Kukliansky, T. Milo, S. Novgorodov, and A. Somech. 2015. "Managing General and Individual Knowledge in Crowd Mining Applications." In *CIDR*.

Crall, A., G. Newman, C. Jarnevich, T. Stohlgren, D. Waller, and J. Graham. 2010. "Improving and Integrating Data on Invasive Species Collected by Citizen Scientists." *Biological Invasions* 12 (10): 3419–28.

Franklin, M., D. Kossmann, T. Kraska, S. Ramesh, and Reynold Xin. 2011. "CrowdDB: Answering Queries with Crowdsourcing." In *ACM SIGMOD*, 61–72.

Lasecki, W., L. Rello, and J. Bigham. 2015. "Measuring text simplification with the crowd." In *Web for All Conference*.

Lukyanenko, R., J. Parsons, and Y. Wiersma. 2014. "The IQ of the Crowd: Understanding and Improving Information Quality in Structured User-Generated Content." *Information Systems Research* 25 (4): 669–89.

Park, H., R. Pang, A. Parameswaran, H. Garcia-Molina, N. Polyzotis, and J. Widom. 2013. "An Overview of the Deco System: Data Model and Query Language; Query Processing and Optimization." *SIGMOD Record* 41 (4): 22–27.

Sheng, V., F. Provost and P. Ipeirotis. 2008. "Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers" In: *ACM SIG KDD*, 614–222.