Mob Data Sourcing

Daniel Deutch
Ben Gurion University of the Negev
deutchd@cs.bgu.ac.il

Tova Milo
Tel Aviv University
milo@cs.tau.ac.il

ABSTRACT

Crowdsourcing is an emerging paradigm that harnesses a mass of users to perform various types of tasks. We focus in this tutorial on a particular form of crowdsourcing, namely crowd (or mob) *datasourcing* whose goal is to obtain, aggregate or process data. We overview crowd datasourcing solutions in various contexts, explain the need for a *principled solution*, describe advances towards achieving such a solution, and highlight remaining gaps.

Categories and Subject Descriptors

H.2.4 [**Systems**]: [Relational Databases]

General Terms

Algorithms, Languages, Design

Keywords

Crowdsourcing, Declarative Systems

1. INTRODUCTION

Crowdsourcing is a powerful new project management and procurement strategy that enables the realization of values associated with an 'open call' to an unlimited pool of people, typically through Web-based technology [6, 28, 18, 26, 22]. We focus on an important form of crowdsourcing where the crowd's task is to generate or 'source' **data**. Generally speaking, crowd-based data sourcing is invoked to obtain data, to aggregate and/or fuse data, to process data, or, more directly, to develop dedicated applications or solutions over the sourced data. With the popularity of the Web, we are increasingly overwhelmed by the quantity of data that is published. Crowd data sourcing brings to light, out of the huge, inconsistent and unverified Web ocean, an important body of knowledge that would otherwise not be attainable. Crowd-based data sourcing democratizes data-collection, cutting companies and researchers reliance on stagnant, overused datasets and bears great potential for revolutionizing our information world.

Wikipedia [30] is probably the earliest and best known example of crowd-sourced data and an illustration of what can be achieved with a crowd-based data sourcing model. Other examples include social tagging systems for images - which harness millions of Web users to build searchable databases of tagged images - traffic information aggregators like *Waze* [29] and hotel and movie ratings like *TripAdvisor* [27] and *IMDb* [16].

However, fulfilment of the great potential in crowd data sourcing has been limited to only a handful of successful projects such as those listed above. This comes notably from the difficulty of managing huge volumes of data and users of questionable quality and reliability. Every single initiative had to battle, almost from scratch, the same non-trivial challenges. The ad hoc solutions, even when successful, are application specific and rarely sharable.

This calls for a principled solution, that will allow to realize crowd data sourcing more effectively and automatically, be able to reuse solutions, and thereby to accelerate the pace of practical adoption of this new technology that is revolutionizing our life.

The development of such a principled solution is a grand goal of research on crowd datasourcing and is the focus of this tutorial.

We believe that database researchers are particularly well-equipped to study the design of such a principled solution. To understand why (and how), it is important to examine the significant conceptual and technical challenges that need to be addressed. The first challenge is utilizing the collected data effectively for answering queries of interest. This challenge stems from the fact that data supplied by the crowd may be erroneous or contradictory. Furthermore, a crowd may be unintentionally slanted or imbalanced with respect to general project-related philosophies, so offering a disproportionate perspective and results. Query answering thus involves the identification of correct and valuable contributions and further providing to users explanation/justification for why these answers are estimated to be correct. A second challenge is identifying what kind of user input would be helpful and which users could be asked to supply it. This challenge involves knowing which of the already contributed data pieces require validation, which pieces are still completely missing, and which users and contributions are likely to be more reliable/enriching. (Incentives are also an issue here). A key difficulty here lies in the recursive dependency between these two challenges: to motivate relevant users to

SIGMOD '12, May 20–24, 2012, Scottsdale, Arizona, USA.

contribute in a helpful manner, we must know which data pieces may be incorrect; but, to determine this, we need to know whether the contributing users are trustworthy, this in turn depending on the estimated correctness of their contributions.

Scaling is particularly important here. We are accustomed to a world with billions of Web pages; we must now get accustomed to a world in which the number of facts is counted in teras. Given the large volumes of data involved, there is no way to have a human verify and correct crowd-sourced data; rather, we must develop automated processes that do the bulk, if not all, of the work. Finally, being set in a generally decentralized Web environment, data sourcing comes with all the common difficulties of large scale Web-data integration: assembly of information that comes from different groups, in different formats, with different first languages and different cultures. Note however that, while we would of course want to make the data coherently integrated, the "noise" (seemingly inconsistent data) itself may be valuable as it may include new unforeseen information. Contradictory input here is thus both a disadvantage and a goal.

Database research has addressed such challenges related to uncertainty, conflict resolutions, trust, recursive relationships, scaling, data integration, etc.,through research on probabilistic databases (e.g. [17, 4, 5, 10, 25]), provenance (e.g. [15, 14, 7, 9, 8]), corroboration (e.g. [12, 21], Datalog (see e.g. [3, 24]) and Web-based data management and integration (e.g. [2, 1, 19, 13]). Adapting the developed technology and solutions to the context of crowd datasourcing is a non-trivial task, but it is in our opinion a promising research area with great potential for breakthrough results.

In this tutorial, we will describe different existing (and missing) techniques for crowd data sourcing, in different contexts. Keeping in mind the grand goal of finding a principled solution, we will identify the similarities and differences between those techniques, and in particular we will pinpoint the common patterns in these solutions, that could serve as desiderata of a unifying model. We will describe recent attempts towards such a principled solution, and discuss them in light of the desiderata, highlighting some gaps. Finally, we will describe recent advancements in database research (in particular [23, 11, 20]) that can be utilized for this effort, discuss to what extent they account for the needs arising in the different contexts, and identify the remaining gaps and open problems. We will do this with illustrative examples, relevant to data management researchers as well as practitioners. Our tutorial will depict a unifying picture of the topic, thus allowing the audience a better understanding of the different approaches, needs, and advancements towards a principled solution. We expect that the tutorial will encourage and guide research on this important area.

2. TUTORIAL OUTLINE

We next describe the main topics that will be discussed in the tutorial.

2.1 Crowd DataSourcing

We will start the tutorial by describing the general paradigm of Crowdsourcing, that harnesses a mass of users to perform various types of complicated tasks. In particular, we will overview the use of crowdsourcing for tasks such as object recognition in images the collection of user preferences, improving the quality of search engines and completing missing information in social networks, such as tags associated with its members.

We will then focus on a particular type of crowdsourcing, namely crowd *datasourcing*, that aim to use the collective wisdom to construct a large database of facts. In particular, we will describe the use of *games* that has emerged as a tool for crowd datasourcing. We will introduce the tasks that crowd datasourcing typically addresses, and the proposed solutions. We will identify common and distinguishing features of datasourcing with respect to general crowdsourcing.

2.2 Towards a principled solution

Much research has been recently directed in the databases community to the development of DB platforms that allow for declarative specification of the crowdsourced data components. These platforms are providing declarative language support and tools to define what data will be retrieved from the crowd (e.g. the choice of questions to ask the crowd).

In this part we will discuss the need for a declarative, principled solution, the advances towards such a solution, and the remaining gaps. We will start by "compiling" the common features of crowd datasourcing into desiderata of a principled solution. We will explain the potential benefits We will present the recently developed declarative tools and techniques that propose partial solution to the problem, and identify gaps between the desiderata and the state-of-the-art. In particular, we will highlight the need for supporting uncertainty, provenance and recursive deduction as well as effective means for corroborating conflicting facts.

2.3 Harnessing existing techniques

As mentioned above, there have been major advancements towards declarative solutions for crowd datasourcing. We claim that such solutions can be enhanced by employing common techniques that were developed in other branches of database research. To this end, we will briefly review the state-of-the-art in a number of relevant areas including management of probabilistic databases, provenance, corroboration, Datalog, and scalable Web-based data management and integration. We will highlight the potential of employing these techniques in the context of crowd data sourcing, as well as the difficulties in this respect; and we will present recent developments in this vein.

Acknowledgments

This work has been partially supported by the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement 291071-MoDaS, by the Israel Ministry of Science, and by the Binational (US-Israel) Science Foundation.

3. REFERENCES

[1] S. Abiteboul, O. Benjelloun, and T. Milo. The active xml project: an overview. *VLDB J.*, 17(5), 2008.

[2] S. Abiteboul, M. Bienvenu, A. Galland, and E. Antoine. A rule-based language for web data management. In *PODS*, 2011.

[3] S. Abiteboul, R. Hull, and V. Vianu. *Foundations of Databases*. Addison-Wesley, 1995.

[4] L. Antova, T. Jansen, C. Koch, and D. Olteanu. "Fast and Simple Relational Processing of Uncertain Data". In *Proc. ICDE*, 2008.

[5] O. Benjelloun, A. D. Sarma, C. Hayworth, and J. Widom. An introduction to uldbs and the trio system. *IEEE Data Eng. Bull.*, 29(1):5–16, 2006.

[6] D. C. Brabham. Crowdsourcing as a Model for Problem Solving: An Introduction and Cases. *Convergence*, 14(1):75–90, 2008.

[7] P. Buneman, J. Cheney, and S. Vansummeren. On the expressiveness of implicit provenance in query and update languages. *ACM Trans. Database Syst.*, 33(4), 2008.

[8] P. Buneman, S. Khanna, and W. C. Tan. Why and where: A characterization of data provenance. In *Proc. of ICDT*, 2001.

[9] J. Cheney, S. Chong, N. Foster, M. I. Seltzer, and S. Vansummeren. Provenance: a future history. In *Proc. of OOPSLA*, 2009.

[10] D. Deutch, C. Koch, and T. Milo. On probabilistic fixpoint and markov chain query languages. In *PODS '10*.

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

[12] A. Galland, S. Abiteboul, A. Marian, and P. Senellart. Corroborating information from disagreeing views. In *WSDM '10*.

[13] L. Gravano, P. G. Ipeirotis, N. Koudas, and D. Srivastava. Text joins in an rdbms for web data integration. In *Proceedings of the 12th international conference on World Wide Web*, WWW '03, pages 90–101, New York, NY, USA, 2003. ACM.

[14] T. J. Green, G. Karvounarakis, Z. G. Ives, and V. Tannen. Update exchange with mappings and provenance. In *Proc. of VLDB*, 2007.

[15] T. J. Green, G. Karvounarakis, and V. Tannen. Provenance semirings. In *Proc. of PODS*, 2007.

[16] Imdb. http://www.imdb.com/.

[17] R. Jampani, F. Xu, M. Wu, L. L. Perez, C. Jermaine, and P. J. Haas. Mcdb: a monte carlo approach to managing uncertain data. In *SIGMOD '08*.

[18] H. Ma, R. Chandrasekar, C. Quirk, and A. Gupta. Improving search engines using human computation games. In *CIKM '09*.

[19] J. Madhavan, S. R. Jeffery, S. Cohen, X. (luna Dong, D. Ko, C. Yu, A. Halevy, and G. Inc. Web-scale data integration: You can only afford to pay as you go. In *CIDR*, 2007.

[20] A. Marcus, E. Wu, S. Madden, and R. C. Miller. Crowdsourced databases: Query processing with people. In *CIDR*, pages 211–214, 2011.

[21] A. Marian and M. Wu. Corroborating information from web sources. *IEEE Data Eng. Bull.*, 34(3):11–17, 2011.

[22] Amazon's mechanical turk. https://www.mturk.com/.

[23] A. Parameswaran, A. D. Sarma, H. G.-M. andNeoklis Polyzotis, and J. Widom. Human-assisted graph search: It's okay to ask questions. In *VLDB*, 2011.

[24] R. Ramakrishnan and J. D. Ullman. A survey of research on deductive database systems. *Journal of Logic Programming*, 1993.

[25] J. Stoyanovich, S. Davidson, T. Milo, and V. Tannen. Deriving probabilistic databases with inference ensembles. In *To appear in Proc. of ICDE*, 2011.

[26] Top coder. http://www.topcoder.com/.

[27] Tripadvisor. http://www.tripadvisor.com/.

[28] L. von Ahn and L. Dabbish. Designing games with a purpose. *Commun. ACM*, 51(8):58–67, 2008.

[29] Waze. http://www.waze.com/.

[30] Wikiepdia. http://www.wikipedia.org/.