Crowdsourcing High-Quality Structured Data

Harry Halpin and Ioanna Lykourentzou

[1] World Wide Web Consortium/MIT, 32 Vassar Street Cambridge, MA 02139 USA
hhalpin@w3.org,
[2] CRP Henri Tudor, Luxembourg ioanna.lykourentzou@tudor.lu

Abstract. One of the most difficult problems faced by consumers of semi-structured and structured data on the Web is how to discover or create the data they need. On the other hand, the producers of Web data do not have any (semi)automated way to align their data production with consumer needs. In this paper we formalize the problem of a *data marketplace*, hypothesize that one can quantify the value of semi-structured and structured data given a set of consumers, and that this quantification can be applied on both existing data-sets and data-sets that need to be created. Furthermore, we provide an algorithm for showing how the production of this data can be crowd-sourced while assuring the consumer a certain level of quality. Using real-world empirical data collected via data producers and consumers, we simulate a crowd-sourced data marketplace with quality guarantees.

Keywords: crowdsourcing; structured data; resource allocation; human computation

1 Introduction

Given there are few things more valuable in the information economy than having the data that matches one's need, we find it likely that accurate and well-maintained data has a monetary value, even though currently most efforts on producing structured data for the Web have been so far focusing on public open data that is created at public cost and published for anyone to use. Therefore, our first hypothesis is that that consumers of data can use *financial incentives* to attract domain experts to produce and update structured data. In this way, the lack of structured data could be corrected by a *data marketplace*, a service that matches the consumers of structured data to the data producers.

To test our hypothesis, we model three data marketplace systems: i)a "Non-profit" system, which functions similarly to the CKAN community, a data production user community without monetary incentives. This serves as a baseline model for a data marketplace without any financial incentives (i.e. producers select data-sets based based on their interests, which we model as a random distribution of expertise), a ii) "Simple profit" system, where producers select micro-tasks based only on their expected profit and a iii) "Smart profit" system, which features a smart scheduling algorithm where producers select micro-tasks among the recommendations of the algorithm given in Figure 1.

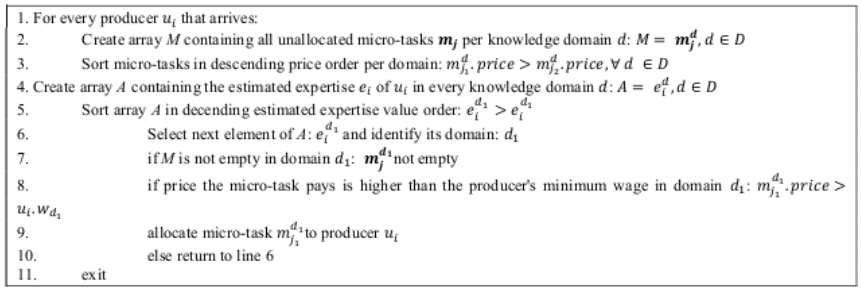

Fig. 1. Smart profit directed crowd-sourcing algorithm

2 Empirical Parameterization

In order to make our simulation realistic, we empirically determined our parameters from The Data Hub, a website that features a 4,826 public datasets.[3] Data-sets are added and kept up-to-date by the voluntary contributions of users, with no financial incentives in place. In order to be concise, we abbreviate The Data Hub as CKAN, given that it was built using the Comprehensive Knowledge Archive Network (CKAN). The site has been crowd-sourcing contributions of structured data-sets since June 2006 and is probably the largest source of open public structured data-sets on the Internet.[4] The website features vastly different sizes and kinds of data-sets, such as Canada's Open Government data, bibliographic data, biological data such as BioPortal. Data-sets are given domains in a "bottom-up" fashion by tagging. The data-sets come in a variety of formats ranging from XML to RDF to CSV. User contributions and modifications of data-sets are recorded and available as metadata for each data-set. Currently, the Data Hub has 3,700 users, although only 1,654 have actually contributed to data-sets. On November 18th 2012, we crawled and used the CKAN API to get statistics for all CKAN data-sets, including their number of tags, creation date, revisions, and users. We use the data as the basis of the parameterization of our simulation. The dataset extracted from CKAN covers a timespan of 67 months, and so we set the simulation time equal to 67 simulation units, with each unit being equivalent to one month, so that we can compare accurately to an empirical baseline, the CKAN data-set. The rest of the extracted parameters used to run the simulation are depicted in Table 2.

3 Results

Our simulation was run using the empirically-derived and estimated parameters given in Table 2, and the results are examined in detail to test our hypotheses,

[3] *http://datahub.io/.*

[4] While Infochimps claims to be larger (approximately 9,000 data-sets) than The Data Hub, approximately half of its data-sets are APIs rather than structured data and thus cannot be queried, and no history of users and revisions are available as Infochimps does not use crowd-sourcing.

Name	Value
Simulation time	67
Users	1654
Domains	20
Micro-tasks per job	3
Dataset inter-arrival time	$\frac{1}{\lambda} = \frac{1}{g'(t)} = \frac{1}{\gamma \cdot \sigma \cdot e^{\sigma \cdot t}}$, with $\gamma = 139.1$, $\sigma = 0.05328$
User inter-arrival time	$\frac{1}{\lambda} = \frac{1}{f'(t)} = \frac{1}{\alpha \cdot \beta \cdot e^{\beta t}}$, with $\alpha = 34.01$, $\beta = 0.05898$
User expertise distribution	normal($\bar{x} = 0.5, \sigma = 0.3$)
Cost distribution	beta([1000, 10000])

Table 1. Parameters of model

namely that financial incentives will increase the quality of dataset production and that a crowd-sourcing co-ordination mechanism (formulated and solved by us as a resource-scheduling problem) will increase the quality of dataset production.

The performance of the three algorithms in terms of average dataset quality (the maximization of which is the objective of the resource scheduling algorithm) is given in Fig. 2 by a quality histogram of the produced datasets. In this diagram, the x axis represents the quality in the range [0,10], with 0 representing no quality at all and 10 the highest possible quality. The figure has two y axes, with the left corresponding to the datasets produced by the "simple profit" algorithm and the right axis illustrating the datasets produced by the "non-profit" and the "smart profit" ones. As can be observed in Fig. 2, the smart scheduling algorithm achieves much higher quality compared to both the non-monetary and the simple profit algorithms. This can be attributed to the fact that the algorithm identifies the domain each user is mostly expert at and "guides" the user's contribution towards a micro-task from this domain. It is also straightforward to observe that the simple profit system (egoistic-only) performs very poorly in terms of quality, because users in this system have no incentive to make good contributions and they rather select the micro-task they will undertake based only on the price it pays. Interestingly enough, perhaps the above rationale can be used to partially explain the low quality results for which purely-profit based systems such as Amazon Mechanical Turk have been often criticized for.

Fig. 3 compares the three algorithms in terms of production efficiency, defined in terms of the ratio of completed and partially completed datasets. The term "completed" refers to the datasets whose micro-tasks have all been allocated and completed by producers and the term "partially completed" refers to the datasets with one or more completed micro-task that have not been fully completed. Fig. 4 shows the production process of these datasets in relation to time. The simple profit system produces a similar number of partially completed datasets as the non-profit system. The smart scheduling algorithm does not produce many partially complete datasets as it produces by far mostly completed data-sets. Strangely enough, the non-profit system produces more completed data-sets than the simple profit systems. This is likely because in the non-profit system, the users attempt to self-allocate according to their domain expertise, but do not do it as efficiently when compared to the smart scheduling algorithm. In

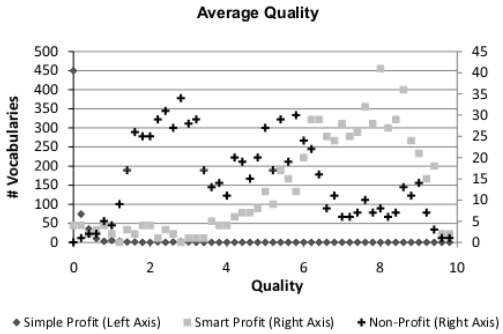

Fig. 2. Quality of Dataset per crowd-sourcing algorithm

contrast, the simple profit system produces many partially completed results of low quality as users are optimizing for profit but ignoring their own domain of expertise. Overall, the contributions of the producers are more focused when the smart scheduling algorithm is used, and they tend to be more dispersed when users self-select the tasks that they will undertake (with this dispersion being higher in the case of the simple profit system due to the over-riding of profit in comparison to domain knowledge).

Therefore, our first hypothesis is is incorrect. *Financial incentives by themselves do not produce higher-quality data-sets, but instead skew the creation of data-sets towards a lower-quality due to optimization of cost over quality by the contributors.* However, our second hypothesis is correct. *The combination of financial incentives combined with a smart scheduling algorithm that directs producers of data-sets to tasks in their domain of expertise produces higher-quality datasets.*

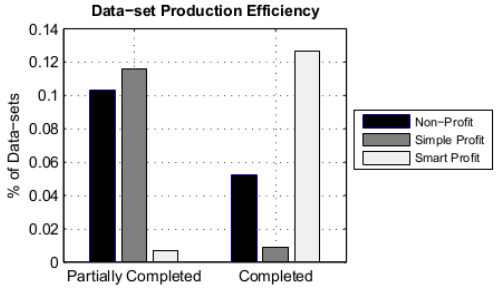

Fig. 3. Dataset completion rates

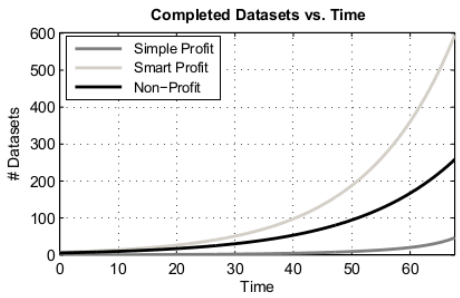

Fig. 4. Completion of Datasets over time

4 Related Works, Conclusion and Future Steps

Although the vision of a data marketplace is not without precedence (e.g. see [4] for an initial vision of the Web as an "Information Marketplace"), much of the data produced to-date is erroneous and not well maintained [6]. Yet, indications

exist that crowdsourcing user expertise can have a positive impact on data quality and evaluations, as shown by paradigms like DBPedia [2], the Datawiki [3] and Google Fusion Tables [5]. Nevertheless, the attempts for financially-powered data marketplaces have so far been very rare, mainly due to the absence of performance guarantees for consumers, and the fact that research on computational methods for gathering data from people in a systematic manner and with performance guarantees is still in its infancy (indicatively see [1, 7]).

In this paper we argue for a Data Marketplace System where various actors can produce and consume datasets for financial contributions. Our first hypothesis was that financial incentives inside such a marketplace would lead to the production of greater amounts of high-quality structured data than systems without such incentives. This hypothesis was shown to be only partially correct: Only with high arrivals of producers into the system does such a simple data market-place succeed. However, here lies an important caveat: our second hypothesis, that crowd-sourcing with financial incentives would perform better in terms of quality if a directed crowd-sourcing algorithm is used (as given in our example by a smart resource scheduling algorithm) was shown to be correct.

These experimental results are only the first foray into the emerging field of crowd-sourcing structured data with performance guarantees. As part of our future work we intend to examine in more detail real data produced by users dynamically over time and to test these against our models. We also plan to examine how non-financial data marketplaces (like TREC, where consumers can access datasets in exchange for expert assessments) could fit and benefit from our model. Last, we hope that this work will trigger a broader discussion related to the use of financial incentives for structured and unstructured web data production and the mechanics of ensuring data production quality in crowdsourcing.

References

1. Yael Amsterdamer, Yael Grossman, Tova Milo, and Pierre Senellart. Crowd mining. In *SIGMOD Conference*, pages 241–252, 2013.
2. S. Auer, C. Bizer, J. Lehmann, G. Kobilarov, R. Cyganiak, and Z. Ives. DBpedia: A nucleus for a web of open data. In *Proc. of the International and Asian Semantic Web Conference (ISWC/ASWC2007)*, pages 718–728, Busan, Korea, 2007.
3. P. Buneman, J. Cheney, S. Lindley, and H. Müller. The database wiki project: a general-purpose platform for data curation and collaboration. *SIGMOD Record*, 40(3):15–20, 2011.
4. M. Dertouzos and B. Gates. *What Will Be: How the New World of Information Will Change Our Lives.* HarperCollins, New York City, United States, 1998.
5. H. Gonzalez, . Y. Halevy, A. Langen, J. Madhavan, R. McChesney, R. Shapley, W. Shen, and J. Goldberg-Kidon. Socialising data with google fusion tables. *IEEE Data Eng. Bull.*, 33(3):25–32, 2010.
6. H. Halpin, P. J. Hayes, J. P. McCusker, D. L. McGuinness, and H. S. Thompson. When owl: sameas isn't the same: an analysis of identity in linked data. In *Proc. of the 9th international semantic web conference on The semantic web*, pages 305–320, Berlin, Heidelberg, 2010. Springer-Verlag.
7. I. Lykourentzou, D. J. Vergados, and Y. Naudet. Improving wiki article quality through crowd coordination: A resource allocation approach. *Int. J. Semantic Web Inf. Syst.*, 9(3):105–125, 2013.