Market-based SPARQL brokerage: Towards Economic Incentives for Linked Data Growth

Mengia Zollinger[1], Cosmin Basca[1], Abraham Bernstein[1]

DDIS, Department of Informatics, University of Zurich, Switzerland
{lastname}@ifi.uzh.ch

Abstract. The growth of the Web of Data (WoD) has primarily been funded by subsidies. New datasets are financed via public funding or research programs. This may eventually limit the growth and could hamper data quality for lack of clear incentives. We propose `MaTriX` , a market-based SPARQL broker over the WoD as an economically viable growth option. Similar to others, queries are associated with a budget and minimal result-set quality. The broker then employs auction mechanisms to find a set of data providers that jointly deliver the results. Preliminary results shows that mixing free and commercial providers exhibits superior: consumer surplus, producer profit, total welfare, and recall.[1]

1 Introduction

A fast growing decentralized repository of knowledge, the WoD consists of many data tenants that publish and manage interlinked RDF datasets. In contrast to the traditional Web the vast majority of datasets are freely available forming the Linked Open Data (LOD) cloud. Most exist by means of subsidies from research projects or governments. Unlike the Web, where the provision of advertising drives much of the monetization, traditional incentive mechanisms to publish data do not work in the WoD, as datasets are algorithmically queried. Consequently, results of WoD queries often do not contain attributions to the original source removing most non-monetary benefit for the publisher. Centralizing all data in one big database akin to Google's index although technically possible would not work as Van Alstyne *et al.* [4] argue since incentive misalignments would lead to enormous data quality problems and inefficiencies, removing this approach from consideration. In this poster we propose *the use of a global price-market for data* to address the questions of economic viability of the WoD. Such a market could close the gap between applications needing access to diverse data and providers whilst providing a well-defined set of incentive mechanisms to all parties. End-users would query the marketplace that would negotiate with providers for answers. Early research on microeconomic-based scheduling focused on the efficiency of computational resources allocation [2]. Stemming from the interaction between data management and microeconomics, the Mariposa [3] RDBMS drops the traditional cost-based optimizer in favor of a market-based one. This poster, in contrast, explores economic methods as incentive mechanisms for publishing.

[1] The poster explains both the procedure in detail and provide the empirical results.

2 System Design

An extension of AVALANCHE [1] `MaTriX` requires that providers register prior to query execution by supplying simple information like the SPARQL endpoint URL and simple statistics like vocabularies used. Then the quality of a provider[2] is assessed. An actual query A with budget B and quality constraint Q is executed in 3 phases as mandated by AVALANCHE . During the *planing* Avalanche generates the ordered list of plans U_A for query A. Then, during *bidding and plan selection* `MaTriX` starts a reverse auction for each of the top k plans $P_i \in U_A$ by issuing requests for bids to the providers who have information pertinent to any query fragment $F_{P_i} \in P_i$. From all successful auctions `MaTriX` picks the plan P_i that minimizes the monetary cost and is still under budget whilst providing sufficient quality (i.e., $i : min_i[\sum(cost(F_{P_i})] < B \wedge min(quality(F_{P_i})) \geq Q$). Finally, `MaTriX` passes the winning plan P_i to AVALANCHE for execution and pays out the fees to the providers.

3 Findings

In a preliminary evaluation we compare settings where data is offered commercially only to settings where there are both free and commercial data providers. Note that we omit the setting with only free providers as it is only viable given subsidies. To generalize our findings we vary the quality of various providers between high, medium, and low as well as vary the kind of auction used between first-price and second price sealed bid auction.

We find that a reverse, sealed-bid second price auction mechanism provides consumers with a higher consumer surplus, in a mixed commercial and free provider setting. Whilst this may not be surprising, we also show that producers will be able to reap higher profits in the mixed setting. We even find, that for profit producers might be enticed by these higher profits to cross-subsidize the free information providers – a surprising result. Furthermore, another positive aspect is denoted by the general incentive for providers to expose high(er) quality data which in turn drives profits up. Whilst our findings are clearly preliminary and burdened by a number of limitations our papers presents the first systematic study trying to provide a sound economic foundation for the WoD. Indeed, the study lays out the foundation and agenda for such economic studies of data on the web. As such it paves the way to a financially healthy WoD.

References

1. C. Basca and A. Bernstein. Avalanche: Putting the Spirit of the Web back into Semantic Web Querying. In *The 6th International Workshop on Scalable Semantic Web Knowledge Base Systems (SSWS2010)*, Nov. 2010.
2. T. W. Malone, R. E. Fikes, and M. T. Howard. Enterprise : a market-like task scheduler for distributed computing environments, Nov. 1983.
3. M. Stonebraker, P. M. Aoki, W. Litwin, A. Pfeffer, A. Sah, J. Sidell, C. Staelin, and A. Yu. Mariposa: a wide-area distributed database system. *The VLDB Journal The International Journal on Very Large Data Bases*, 5(1):48–63, Jan. 1996.
4. M. Van Alstyne, E. Brynjolfsson, and S. Madnick. Why not one big database? principles for data ownership. *Decis. Support Syst.*, 15(4), Dec. 1995.

[2] a procedure beyond the scope of this poster