Data Quality Adjustments for Pricing on Data Marketplaces

Florian Stahl[1] and Gottfried Vossen[1,2]

[1] ERCIS, Leonardo-Campus 3, 48149 Münster, Germany,
{Stahl,Vossen}@ercis.de
[2] University of Waikato Management School,Private Bag 3105,
Hamilton, 3240, New Zealand

Abstract. Currently, information has become an increasingly important production factor which has led to the emergence of data marketplaces that leverage big data technologies. However, value attribution of data is still difficult. This work suggests to discuss what role data quality can play in this context, particularly: what quality measures are relevant in the context of big data, how they can be measured, and how the quality of a data product can be efficiently modified to create different versions[3] of a data product.

1 Introduction

Information has become an important production factor [6]. This has led to a point at which data – as the basic unit in which information is exchanged – is increasingly being traded on data marketplaces, extensively described in [4, 9] and put on the database research agenda by Balazinska et al. [2, 1]. Basically, data marketplaces are platforms levering big data technologies that allow providers and consumers of data and data-related services, such as data mining algorithms, to interact with each other. One prominent German example of such a data marketplace is MIA[4] which employs large computer clusters to crawl substantial parts of the German Web and to provide an analysis infrastructure for the gathered data. This is particularly beneficial for small and medium-sized enterprises as they would otherwise not be able to access and analyse such data. This paper suggests to discuss what role data quality modifications can play in the context of data marketplaces and big data applications building on them.

 In: R. Bergmann, S. Görg, G. Müller (Eds.): Proceedings of the LWA 2015 Workshops: KDML, FGWM, IR, and FGDB. Trier, Germany, 7.-9. October 2015, published at http://ceur-ws.org

[3] In this work, the term *version* will be used in its economic sense, i. e., to refer to different variants of a data product; this is not to be confused with versions as known from temporal databases.

[4] http://mia-marktplatz.de/

2 Pricing on Data Marketplaces

Given that the value for data and data-related services is subjective to its consumers [7], it is not surprising that little sense for its value exists in the database community [2, 1]; this is mainly owing to the fact that data is an information good with peculiarities, such as resemblance to public goods [12].

One approach to reduce the uncertainty for data providers is to apply reverse pricing mechanisms that allow customers to suggest prices, which – if well-designed – allow for a revelation of the customer's true willingness to pay. Reverse pricing has the advantage that customers participate in the pricing process, which is generally seen as positive, even if used for price discrimination – i. e., asking different prices of different customers [3].

Name Your Own Price is such a pricing mechanism which is often employed in auctions, for instance, EBAY'S *make offer* option. In contrast to established physical goods, digital goods, such as data, can be sold multiple times because of the low cost of reproduction. Thus, in order to avoid fierce price wars, it is recommendable to adapt a data product to a customer's preferences, which can also be seen as a further benefit for customers. Although not discussing the economic intuition behind it, TANG ET AL. suggested to use a Name Your Own Price mechanism in the context of data marketplaces. In [10] they suggested to adapt the completeness of XML data and in [11] they focused on the accuracy of relational data based on a customer's bid. Concretely, the provider advertises a price and customers may suggest a price they are willing to pay. If the bid is lower than the ask price, completeness or accuracy of the data product will be lowered to match the offered price. Furthermore, the argument can be made that the threshold can also be hidden from the buyer. In this case the profit increases if the suggested price is higher than the requested price.

3 Discussion

The previously mentioned works focus on only one quality dimension. Therefore, the question remains how this can be adapted to multiple quality criteria, which has been extensively discussed in [8]. As a starting point we suggest to model the distribution of discounts to different quality criteria as a multiple-choice knapsack problem. Given 1) a set of quality criteria, 2) a function that creates versions for all quality criteria, 3) a function that attributes the ask price to these versions, and 4) customer as well as vendor preferences for certain quality criteria, an optimal combination can be calculated even on commodity hardware for a limited number of quality criteria such as those identified by NAUMANN [5].

Having made these calculations, the quality of data products has to be adjusted to match a customer's suggested price. However, at the moment it is not quite clear what data quality dimensions are relevant in the context of big data analysis applications. Thus, a number of questions arise. Consequently, this work suggests to discuss the following questions:

– What quality dimensions are relevant for big data applications?

– How can they be practically applied to large data sets and how can they be measured efficiently?
– And most importantly: how can big data architectures be utilised to adapt the quality of big data products efficiently in order to meet a customer's requirements?

References

[1] M. Balazinska, B. Howe, P. Koutris, D. Suciu, and P. Upadhyaya. "A Discussion on Pricing Relational Data". In: *In Search of Elegance in the Theory and Practice of Computation*. Ed. by V. Tannen, L. Wong, L. Libkin, W. Fan, W.-C. Tan, and M. Fourman. Vol. 8000. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013, pp. 167–173.

[2] M. Balazinska, B. Howe, and D. Suciu. "Data Markets in the Cloud: An Opportunity for the Database Community". In: *PVLDB* 4.12 (2011), pp. 1482–1485.

[3] O. Hinz, I.-H. Hann, and M. Spann. "Price discrimination in e-commerce? An examination of dynamic pricing in name-your-own price markets". In: *MIS quarterly* 35.1 (2011), pp. 81–98.

[4] A. Muschalle, F. Stahl, A. Löser, and G. Vossen. "Pricing Approaches for Data Markets". In: *Proceedings of the Workshop Business Intelligence for the Real Time Enterprise*. Istanbul, Turkey, 2012.

[5] F. Naumann. *Quality-Driven Query Answering for Integrated Information Systems*. Vol. 2261. Lecture Notes in Computer Science. Springer, 2002.

[6] K. North. *Wissensorientierte Unternehmensführung*. 5th edition. Gabler, 2011.

[7] C. Shapiro and H. Varian. *Information Rules: A Strategic Guide to the Network Economy*. Strategy/Technology / Harvard Business School Press. Harvard Business School Press, 1999.

[8] F. Stahl. "High-Quality Web Information Provisioning and Quality-Based Data Pricing". PhD thesis. University of Münster, 2015.

[9] F. Stahl, A. Löser, and G. Vossen. "Preismodelle für Datenmarktplätze". In: *Informatik-Spektrum* 37.1 (2014).

[10] R. Tang, A. Amarilli, P. Senellart, and S. Bressan. "Get a Sample for a Discount". In: *Database and Expert Systems Applications*. Ed. by H. Decker, L. Lhotská, S. Link, M. Spies, and R. R. Wagner. Vol. 8644. Lecture Notes in Computer Science. Springer International Publishing, 2014, pp. 20–34.

[11] R. Tang, H. Wu, Z. Bao, S. Bressan, and P. Valduriez. "The Price Is Right". In: *Database and Expert Systems Applications*. Ed. by H. Decker, L. Lhotská, S. Link, J. Basl, and A. Tjoa. Vol. 8056. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013, pp. 380–394.

[12] L. Vomfell, F. Stahl, F. Schomm, and G. Vossen. *A Classification Framewiork for Data Marketplaces*. Tech. rep. 23. Münster: ERCIS, 2015.