Computers & Industrial Engineering 112 (2017) 1–10

Contents lists available at ScienceDirect

Computers & Industrial Engineering

journal homepage: www.elsevier.com/locate/caie

Data pricing strategy based on data quality ☆

Haifei Yu, Mengxiao Zhang *

School of Business Administration, Northeastern University, Shenyang 110819, PR China

ARTICLE INFO

Article history:
Received 31 December 2016
Received in revised form 2 June 2017
Accepted 7 August 2017
Available online 9 August 2017

Keywords:
Big data
Data marketplace
Data pricing
Production management
Bi-level programming model

ABSTRACT

This paper presents a bi-level mathematical programming model for the data-pricing problem that considers both data quality and data versioning strategies. Data products and data-related services differ from information products or services in terms of quality assessment methods. For this problem, we consider two aspects of data quality: (1) its multidimensionality and (2) the interaction between the dimensions. We designed a multi-version data strategy and propose a data-pricing bi-level programming model based on the data quality to maximize the profit by the owner of the data platform and the utility to consumers. A genetic algorithm was used to solve the model. The numerical solutions for the data-pricing model indicate that the multi-version strategy achieves a better market segmentation and is more profitable and feasible when the multiple dimensions of data quality are considered. These results also provide managerial guidance on data provision and data pricing for platform owners.

1. Introduction

The advent and ubiquity of Web 2.0, social networks, cloud computing, and "Software-as-a-Service" has expanded the volume of personal, business, and public data at an alarming rate. Big data volumes, and the diversity of such data, are a defining feature of the modern world, with significant financial and commercial implications. Enterprises rely not only on the acquisition of data in itself, but also on professional third-party platforms that collect data from various sources (Mohanty, Jagadeesh, & Srivatsa, 2013). Increasingly, data providers appreciate the gradual commercialization of data, and have established network platforms for data trading (Schomm, Stahl, & Vossen, 2013), thereby giving rise to data marketplaces.

Armstrong and Durfee (1998) introduced the term 'data marketplace' to denote the ensemble of agents involved in commercial transactions. A typical data market comprises three main roles: data providers, data consumers, and a data-market owner. Data providers supply data to the data market and set the corresponding prices. Data consumers buy the data that they need. Acting as the intermediary between providers and consumers, the owner negotiates the pricing mechanism with those providers and manages the data transactions (Tang, Amarilli, Senellart, & Bressan, 2014). Currently emerging data platforms include Factual,[1] Infochimps,[2] Xignite,[3] and the Windows Azure Data Marketplace[4] (Stahl, 2013). The latter, for example, encompasses more than one hundred data sources for sale, Infochimps contains 15,000 data collections, and Xignite focuses on financial data.

The emergence of data markets has prompted the design of a new kind of business model in which information and analysis tools effectively become tradable electronic goods (Muschalle, Stahl, Löser, & Vossen, 2012). In data markets, data products are processed and sold like information products at appropriately defined prices to data consumers. The present study defines data products as datasets in the form of tradable data goods after crawling, reformatting, cleaning, encrypting, and other processes. This includes government data, medical data, financial data, e-commerce data, and traffic data.

The pricing of data products is an important issue. Most data-product transactions are completed through offline negotiations between data sellers and buyers, a small proportion of which is done online. The main pricing models for data markets are as follows: (1) Free models are those where data services can be used for free. (2) Freemium models combine free services and value-added services. In the pricing model, consumers have limited access to data for free and pay for the premium services. (3) In packaging models, consumers buy a certain amount of data at a

☆ This work is supported by Fundamental Research Funds for the Central Universities (N150604003).

* Corresponding author.
E-mail address: mengxiaozhang47@hotmail.com (M. Zhang).

[1] https://www.factual.com.
[2] http://www.infochimps.com.
[3] http://www.xignite.com.
[4] https://datamarket.azure.com.

http://dx.doi.org/10.1016/j.cie.2017.08.008
0360-8352/

fixed price. (4) In pay-per-use models, consumers pay for data services based on their usage. (5) Flat-fee models involve data consumers paying a monthly subscription fee in return for unfettered access to data services. (6) In two-part-tariff models, consumers pay a fixed basic fee that becomes supplemented by an additional fee when their usage exceeds some pre-defined quota (Muschalle et al., 2012; Schomm et al., 2013).

Common weaknesses in existing data-pricing mechanisms include (1) the lack of a standardized pricing model. On data platforms such as Aggdata[5] and CustomLists,[6] data are traded mainly through private agreements between data providers and consumers, whereas Infochimps and Azure DataMarket charge their members a monthly subscription fee. (2) Issues relating to data quality tend to be neglected. Few data-pricing models for data markets consider data quality, despite the availability of relevant tools and technologies for assessing and improving data quality. (3) Opaqueness. Pricing strategies are mainly seller-driven, with the cost of data acquisition, cleaning, and packaging being invisible to consumers (Balazinska, Howe, & Suciu, 2011). These shortcomings call for the development of a rigorous and reasonable pricing model for data marketplaces.

The proper assessment of data value is the basis of a rigorous and reasonable data-pricing model. Heckman, Boehmer, Peters, Davaloo, and Kurup (2015) suggest focusing on the intrinsic value and quality of data, instead of the value of the information that underlies the data, in the interest of transparency and fairness. However, data value is determined by many, rather than one, attribute. We therefore consider the multiple dimensions of data quality and establish a linear method of multi-dimensional quality assessment. Data value is also determined by the complex interaction of multiple factors (Heckman et al., 2015). For example, an increase in the timeliness of a particular dataset may occur at the expense of its completeness. Additional costs would therefore be incurred by the data provider to increase the timeliness while simultaneously preserving completeness. By considering the interactions of multiple elements, we establish a nonlinear method for evaluating the integrated value of data.

The present study aims at realizing an effective data valuation on a data-market platform by extending it to an integrated and multi-dimensional quality assessment. Furthermore, we examine whether or not the multi-version strategy is suitable for a data-market owner when considering the linear and the integrated assessment model, and provide some guidance to the data-platform owner on how to produce, provide, and price data products.

Based on the linear and the integrated assessment model, we adopt the perspective of the data value and consider both the profit derived from a data platform and the utility to data consumers, in order to propose a fair and reasonable data-pricing model. We establish a bi-level programing model with two kinds of cost functions to analyze the production-decision behavior of the data-platform owner and the purchasing-decision behavior of data consumers. Data-platform owners may have some monopoly power that allows them to personalize pricing through price segmentation, including versioning, segmenting, and negotiating (Pantelis & Aija, 2013). In the model, as a leader, the owner decides the planned number of data-product versions, the data quality, and the prices accordingly; as followers, consumers choose the ideal data product that is provided on the data platform and that maximizes utility. The model determines the actual number of versions, data qualities, and the corresponding prices based on the total revenue of the data-platform provider and utility of each consumer, and assist consumers and providers in making reasonable decisions. The features of multiple versions are analyzed and managerial implications are presented for data-platform owners.

The rest of this paper is organized as follows: we firstly review the existing relevant literatures in Section 2. Section 3 then describes the data-pricing problem, based on data quality, and establishes a bi-level programing model that involves a data-platform owner and data consumers. Numerical applications and managerial implications are discussed in Sections 4 and 5, respectively, followed by conclusions and proposed avenues for future work in Section 6.

2. Literature review

The value assessment of intangibles such as intellectual products is not a new challenge for entrepreneurs and scholars. The pricing of information products and information services has generated a substantial literature. We here review representative works on these methods, before selectively reviewing research on data pricing.

Information-service markets involve three commonly used pricing schemes: "pure flat-fee" pricing, "pure usage-based" pricing, and "two-part tariff" pricing (Wu & Banker, 2010). Wu and Banker (2010) found that marginal and monitoring costs can influence a firm's choice of pricing scheme. Huang, Kauffman, and Ma (2015) argue for the existence of service interruptions in cloud software, to which some consumers are sensitive. In such a market, it is sensible for a vendor of cloud-computing services to adopt a hybrid pricing strategy that mixes fixed-price reserved services with spot-price on-demand services. Mei, Li, and Nie (2013) constructed a pricing model based on the Stackelberg game and advocated adopting a pure-bundling strategy, instead of pure components, when device prices are high and consumers' evaluations vary widely. Balasubramanian, Bhattacharya, and Krishnan (2015) considered differences in the use of frequencies and the psychological costs to consumers that are associated with a pay-per-use model. They concluded that two factors can affect a seller's profit by analyzing two pricing mechanisms for information products, namely the fixed-fee and pay-per-use mechanisms. Sundararajan (2004) argued that administering usage-based pricing incurs transaction costs, which influence the optimal pricing of information goods when the available information is incomplete.

On the other hand, versioning is a widespread differentiation strategy used in information-product markets. Under this scheme, a firm customizes information products according to the customers' need and encourages them to pay the highest possible price for goods to maximize its overall revenue (Shapiro & Varian, 1998). Bhargava and Choudhary (2001) analyzed the optimal strategy for vertically differentiated information products in the context of a monopoly. They showed that the optimal product line of a firm depends on the benefit-to-cost ratio of qualities when the consumer's valuation is a linear function of product quality and consumer type. Li, Feng, Chen, and Kou (2013) defined a nonlinear function to describe the "willingness to pay" and the utility to a consumer who has a specific quality requirement, and developed hybrid steady-state evolutionary algorithms. They observed that a monopoly can achieve more profit by using a multi-version strategy. Chen and Seshadri (2007) considered a two-stage development problem and found that versioning is an optimal strategy for sellers if the consumers have a convex-shaped reservation utility function. Because data and information products have many features in common, the pricing methods used for information products provide insight for our present research. However, these

[5] http://www.aggdata.com/.
[6] http://www.customlists.net/home/.