2L1-2

データの利用価値を評価するためのモデル構造

Model structure for evaluating the utility value of data

曾　妍媛[*1]　　大澤　幸生[*2]
Yanyuan Zeng　　Yukio Ohsawa

[*1] 東京大学大学院　工学系研究科　システム創成学 1　[*2]（同上）2
Department of Systems Innovation, School of Engineering, The University of Tokyo #1

Since the development of Big Data, we have been doing various researches on data utilization; however not enough attention has been paid to the value of the data itself. Therefore, there is a high risk that the data will not be useful to the user even if developed or mined by massive steps. Additionally, once the value of data cannot be fully demonstrated, it will also cause a great loss in terms of business. As a consequence, in this research, we focus on the present condition that there is no fixed standard in the data value evaluation, the model is structured to evaluate the data from all fields; as well the value of each data has been relatively accurate assessed.

1. はじめに

大規模なデータ集合に対応し、2011 年ごろからビックデータの過剰な流行が始まり[Lohr 12]、そして現在はほぼ終息した。しかし、データの利活用促進に関連する課題と技術を捉える活動は様々に進み、議論されてきた[中野 14, Chen 12]にもかかわらず、人々がデータに価値があることを期待するまでの認知機構の解明は依然として難しい。

データマイニングを通じて、ビジネス上の意思決定や立案に資する度合という意味でのデータ価値の重要性は今までの研究でも論述されている[Chen 12, Berry 97]。データの関連価値を予測するためのメカリズム構造を提唱した論文としては[Allen 74]等から研究されており、変量選択とデータ増加の関連性について議論されている。だが、人が期待するデータの価値が不明なままで利活用の方法の検討、技術的な難点の解決、プライバシーの保護などのプロセスを立ち上げることは本末転倒であろう。ここで考えたいことは、データが利用者にどれほどの利益をもたらすことを人はデータの中身を見ない段階で期待するかである。本研究で明らかにしたいことは、データの利用価値を評価し、意思決定をし、利活用を行う上で、さまざまなデータ利用がどのような認知プロセスを経るかである。

大澤らは、データを秘匿したままで概要情報（Data Jackets:DJ）だけを共有することによって使途を検討するワークショップ IMDJ としてデータ市場を再定義した。ここでのデータ市場は、データ所有者とデータの潜在利用者がデータの価値を評価する行為を通して、新たなデータ利活用方法が生み、新たなビジネスを創成する場である[Ohsawa 13]。また、そのデータ市場をサポートする管理、保存、検索システムも実装され日々改善されている[早矢仕 16]。しかし、IMDJ では実際の金銭でデータや利活用案を購入することにすると、情報の不透明性によってアイデアが生まれにくくなる。そこで、仮想的な紙幣（モンキー）によってデータ市場での取引対象を評価する、模擬市場を構成することが多い。しかし、DJ と仮想紙幣を用いて模擬売買されるときのデータの価格は、実際にデータを取引する段階での値段とは必ずしも一致しない。データの価値は様々な原因、要素から構成されることを考慮すれば、更に詳細な評価方法が求められる。

連絡先：曾妍媛，東京大学大学院，東京都文京区弥生 2-8-2，07040179173，zengyy1992@yahoo.com

<DJ15: Market share data>	
Outline	Vehicle registration or sales data is used to grasp company's positioning in the market . This data is one of competitive environment information with regard to heady duty, midiium duty and light duty trucks in Asia.
Variables・Attributes	market, brand, time, product, number of sales or registration
Sharing Policy	Under particular condition.
Types	Time Series, Table
Formats	CSV, XLS, etc.

図 1 データジャケットの一例

2. モデルの構造

本論文では、大澤研究室が独自開発した DJ 検索システム（Data Jacket Store, 以下は DJ store ）[早矢仕 16]を基に、全体の評価システムを構想している。図 2 は構想するシステムの概要図である。

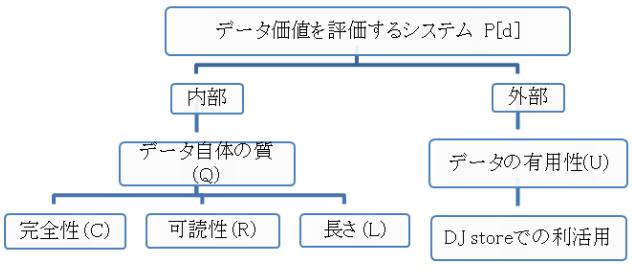


図 2 全体評価システム概要図

データを評価する際に、このシステムでは二つの枠で分析を行う。まず、データ自体を分析する枠である（図 1 では内部という）。そこで、データの質(以下頭文字 Q で表示)を三つの側面から評価する。一つ目は DJ の完全性(DJ に記入されている度合で、以下頭文字 C で表示):データ源の信頼性と特徴記述の完全性という視点から評価する。二つ目は DJ の可読性(以下頭

文 R で表示):テキストデータのみについて評価される。三つ目は DJ の長さ(以下頭文字 L で表示):異なる形の DJ は読者にとって一番適切な長さが違う–例えば、テキストは読みやすいので多少長くとも人は読むのであるが、数字の羅列であれば短くなければ人出では処理できず機械に処理を任せることになる。

次に、データがもし与えられた場合、どれくらいの価値を持つとユーザが期待するかを評価する。ここでは、データの有用性というコンセプト(以下頭文字 U で表示)に注目する。DJ Store における利活用程度を基づいてデータの有用性を評価するため、両タイプのモデルを提案する(記号の意味は表1参照)

データ d の中身が得られる場合:

$$P_{D'}[d] = \alpha[Q] + U + \varepsilon(\text{未知潜在価値}) \cdots\cdots ①$$

データなしの場合

$$P[d] = \alpha[Q] \quad \cdots\cdots ②$$

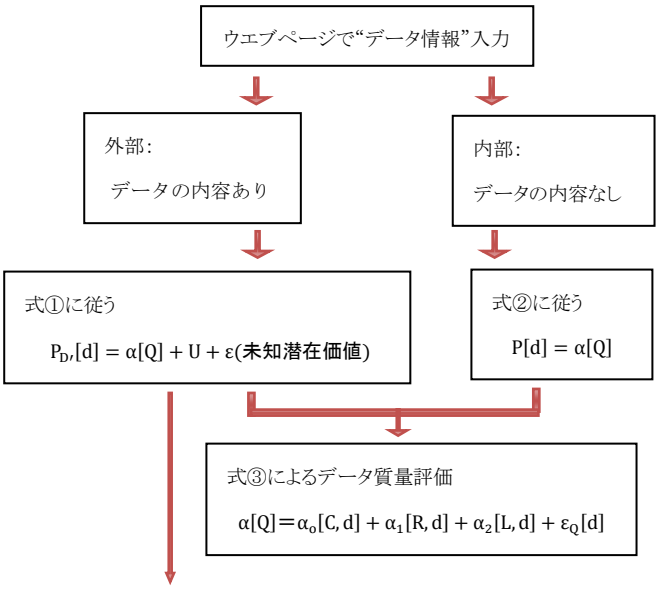
更に、データそのものに期待される品質を式③で定義した:

$$\alpha[Q] = \alpha_o[C,d] + \alpha_1[R,d] + \alpha_2[L,d] + \varepsilon_Q[d] \cdots ③$$

記号	説明
$P_{D'}[d]$	特定なデータに期待されるデータの価値
$P[d]$	データの内容なしで(DJ のみから)期待されるデータの価値
$\alpha[Q]$	期待されるデータ自体の質
U	データの内容の(あるいは期待される)有用性
ε	データの未知潜在価値
$\alpha_o[C,d]$	DJ の完全性
$\alpha_1[R,d]$	DJ の可読性
$\alpha_2[L,d]$	DJ の長さ
$\varepsilon_Q[d]$	誤差

表 1 計算式に使用される記号及び説明

これらの式は、実際にデータの価値評価システムを想定するとすれば、図 3 のような流れで分析することを意味している。データの有用性は本来、特定なデータがあって初めて評価されるため、DJ だけから推定される有用性の期待値 U は DJ の特徴に応じて定められる。この論文では、DJ が DJ Store で使われる回数、また実験型のワークショップでの評判によって U を評価する。



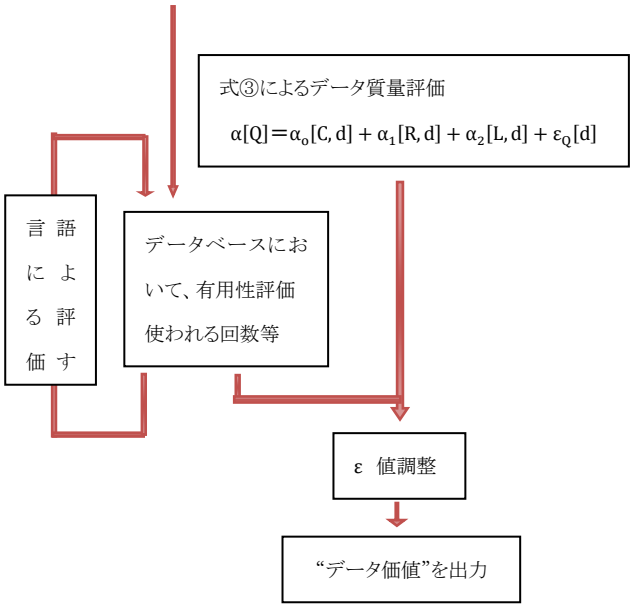


図 3 システム・オペレーティング

3. DJ Store における実験と評価

3.1 変数の設定

独自開発したデータダイジェスト検索システム(DJ Store[早矢仕 16])を利用し、各 DJ の三つの特徴を洗い出した。一つ目データベースにおける DJ のクリックされた回数、二つ目はイノベーションデータ市場(Innovators Marketplace on Data Jackets) [Ohsawa 15]で使われた回数、三つ目は DJ の質である。三つ目のデータの質に関しては、式③にしたがい DJ の完全性、可読性、長さ、全面性から評価する。一つ目と二つ目のパラメータの相関関係は表 2 のように検定された。ここで、度数すなわち DJ の数は、現在 1500 件の DJ のうち両パラメータがいずれも上位にあった 48 件としている。

相関分析			
		DJclick	IMDJ
all_DJclick	Pearson の相関係数	1	.089
	有意確率 (両側)		.549
all_IMDJ	Pearson の相関係数	.089	1
	有意確率 (両側)	.549	

表 2 相関関係検定結果

一つ目の変数と二つ目の変数は相関係数は小さく、U などを評価する上で、片方に縮約することはできない。次に、三つ目のデータ質がその二つの変数といかなる関係を持つかについて検証する。

3.2 データジャケットに基づくデータの質の検証

DJ Store[早矢仕 16]に入っているデータジャケットの例をの一例は図 1 を参考する。

そこで、DJ Store に入っている、クリックされた回数とイノベーションデータ市場に使われた回数とも五回以上のデータを合計 48 件の DJ を対象として実験した。実験方法はデータジャケットの各項目の空欄状況をチェックした。(上記の図 1 において左側は項目で、右側はその説明である。) 単純集計で、空欄であれば、点数をプラスゼロ、そうでなければ点数に 1 を加えた。

Outline, Variables, Sharing Policy, Types, Formats の五項目の点数は Outline score, Variables score, Sharing Policy score, Types score, Formats score とし、最後に合計点数 score を求めた。これらの各項目がそれぞれ、実際にデータのクリック回数とデータ市場で使われた回数には影響を与えている度合を検証するためである。

まず、合計点数 score と click(クリックされた回数)と IMDJ(データ市場で使われた回数)の関係性を検証した。その結果以下の表のようである 。

	click	IMDJ
ScorePearson の相関係数	0.380	−0.079
有意確率 (両側)	0.000	0.000

score と click の間に低相関関係が見られる。

更に、データ質のどの部分は一番読者の選択に影響を与えるかを検証した。5項目とそれぞれclickとIMDJとの相関を表に示す。coreの中の項目Variables scoreはclick(クリックされた回数)との関係性が見いだされる。一方、Outline scoreはclickと関係性があまり高くない。このことから、ユーザがDJをクリックする時、該当するデータの価値はOutline(概要)よりもVariables(変数)に着目して評価されることが分かる。すなわち、ユーザにデータの価値を伝えるためには、データ概要の内容を増やしたり変えたりするよりも、データ変数の説明を充実させる方が効果的であると考えられる。

	click	IMDJ
OutlineScore Pearson の相関係数	0.245	0.109
有意確率 (両側)	0.327	0.462
Variables/AttributesScorePearson の相関係数	0.347	0.057
有意確率 (両側)	0.091	0.703
SharingPolicyScore Pearson の相関係数	−0.60	−0.294
有意確率 (両側)	0.683	0.043
TypesScore Pearson の相関係数	0.022	0.154
有意確率 (両側)	0.880	0.294
FormatsScore Pearson の相関係数	−0.036	−0.097
有意確率 (両側)	0.807	0.511

表 3 5 項目と click と IMDJ の関係性

4. 結論

本論文では、データ市場技術である IMDJ における、すなわち DJ のみを見た場合のデータの期待価値を判断するため、全体の評価システムの構造を提案した。その上で、具体的変数設定とデータ質は他の変数にはどれくらいの影響があるかを検証した。データ評価には三つの変数を提案し、それぞれの重要性にも説明した。

また、DJ の各成分がデータの期待価値に影響する度合についても検証した結果、データ質の変数の影響が一番大きいことを見出した。この点は、データ市場に今後の改善方向を示唆するものである。すなわち、ユーザは抽象的なデータの概要ではなく具体的な変数に注目してデータの利用価値を見積もることになる。

今後の研究として、更にシステムの全体構造を深掘りし、それぞれのブランチでより正確な実行方法の構造を進めていく。

参考文献

[Lohr 12] Steve Lohr: The Age of Big Data, The New York Times ,Feb 11 2012.

[中野 14] 中野美由紀: ビックデータ統合利活用における課題と技術, 電子情報通信学会誌, 2014.

[Chen 12] Hsinchun Chen, Roger H.L. Chiang, Veda C. Storey: Business Intelligence and Analytics: From Big Data to Big Impact, Mis Quarterly, University of Minnesota, 2012.

[Berry 97] Michael J. Berry, Gordon Linoff: Data Mining Techniques: For Marketing, Sales, and Customer Support, John Wiley & Sons, Inc, 1997.

[Ohsawa 13] Yukio Ohsawa,Hiroyuki Kido, Teruaki Hayashi, Chang Liu: Data Jackets for Synthesizing Values in the Market of Data, 17th International Conference in Knowledge Based and Intelligent Information and Engineering Systems, Procedia Computer Science, 2013.

[早矢仕 16] 早矢仕晃章, 大澤幸生: Data Jacket Store:データ利活用知識構造化と検索システム, 人工知能学会論文誌, 2016.

[Ohsawa 15]Yukio Ohsawa , Hiroyuki Kido, Teruaki Hayashi, Chang Liu, Kazuhiro Komoda: Innovators Marketplace on Data Jackets, for Valuating, Sharing, and Synthesizing Data, Knowledge-Based Information Systems in Practice, Springer International Publishing, 2015.

[Allen 74] DavidM. Allen: The relationship between variable selection and data agumentation and a method for prediction, Technometrics, 1974.