1J3-OS-18b-3

データ利活用知識検索システム DJ Store を用いた 利用価値のあるデータの特徴抽出

Feature Extraction of Valuable Data Using Data Utilization Knowledge Retrieval System DJ Store

早矢仕 晃章[*1]
Teruaki Hayashi

大澤 幸生[*1]
Yukio Ohsawa

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

The potential expectation about generating innovative businesses and creating values by combining data from different regions has been increased. In order to lead data-driven innovations, a market of data is expected to enhance this combination via data exchange. Innovators Marketplace on Data Jackets (IMDJ) is a gamified workshop for discovering the value of existent data by discussing to combine data. A Data Jacket is meta-data, summarizing a dataset. Even if the data is not open in public, Data Jacket enables the participants to consider the latent value of datasets through creative communication among stakeholders. Data Jacket Store is a system for retrieving knowledge of data utilization by structuring and reusing information of IMDJ using RDF. In this study, we investigate a feature of valuable data for creating solutions, sharable data, e.g., Open Data or hidden data which is hardly shared in public. The result shows that the data which is hardly open may be recognized valuable for solving problems and creating new businesses.

1. はじめに

1.1 データ利活用に対する期待

ビッグデータという言葉の流行の背景には、蓄積されるデータそのものの増加だけでなく、データから意思決定において重要な価値ある事象や傾向を発見したいという企業や組織の期待があると考えられる。スマートフォンやウェアラブル機器などの個人用端末の普及により、今まで取得困難と言われてきた個人の購買履歴や移動歴などのパーソナルデータが取得可能となってきた。また、製造業の現場では、センサーなどの機器の高度化に伴い、高粒度かつ膨大なデータが取得できるようになった。さらに、オープンガバメントへの取り組みの一環から、データを二次利用可能な形で公開し、再利用性を高めるオープンデータが普及してきた。

これらの変化により、保有するデータを用いたり、他の組織のデータを組み合わせることで新規ビジネスの創出や既存の事業の付加価値向上への潜在的な期待が高まってきていると言える。

1.2 データ市場

データ市場とは、データ利活用によるイノベーションの場及びプラットフォームを意味する[大澤 14]。データ市場では、データの公開・共有が強制されるのではなく，自由市場の原理から利用者が必要なデータを選び，所有者との交渉の末に入手し、データに基づく意思決定やイノベーションの創出が行われる。

データ市場は、データについての情報を Web 上に列挙しただけのものではなく、データ保有者とデータ利用者、利活用方法の提案者のコミュニケーションからデータの価値を発見し、データを適切な価格で売買または交換が行われる場である必要がある。Microsoft Azure Marketplace[*1]、KDnuggets[*2]、Qlik[*3] など、すでに売手と買手が同意した価格でデータやツールを売買するサービスやプラットフォームは存在するものの、データに関わるステークホルダーによる「利用方法の提案」,「評価」というコミュニケーションが行われるプラットフォームとしての機能を有するに至っていない。データに基づくイノベーションの場としてのデータ市場を活性化させるためには、価値あるデータを選んで入手するために必要なだけの交渉や熟考ができる場が必要である.

2. オープンデータと“クローズドデータ”

2.1 データの共有可能性

世の中には様々な形式で保存されたデータが存在するが、これらは 2 種類に大別できると考えられる。オープンデータなどの一般に共有可能なデータと、売買や交渉により共有される可能性のある秘匿データ（クローズドデータ）である。共有可能データは Web 上に公開されていたり、情報開示を求めれば必要に応じて入手可能な状態にあるデータを意味する。一方で、共有できないデータとは、売買や交渉が必要であったり、公開によるリスクを考慮して共有されないデータを指す。

国や地方自治体では，データ利用の制限を緩和し，ビジネスや新規サービスのために役立てる，オープンデータの取り組みが盛んに行われてきている. Linked Open Data（LOD）研究では，RDF（Resource Description Framework）とクエリ言語である SPARQL を用い，行政のオープンデータや Web 上に公開されているデータに統一的にアクセスできる環境の構築を進めている[大向 13].

一方で、ビジネス機会の損失やセキュリティ，個人の識別性の問題などから，企業や個人などのデータは基本的に公開され

連絡先：
早矢仕晃章，東京大学大学院　工学系研究科　システム創成学専攻，teru-h.884@nifty.com
大澤幸生,東京大学大学院　工学系研究科　システム創成学専攻，ohsawa@sys.t.u-tokyo.ac.jp

本研究は JST，CREST の一部です。本研究を支援してくださった構造計画研究所（KKE）の皆様には感謝申し上げます。

*1 https://datamarket.azure.com/
*2 http://www.kdnuggets.com/
*3 http://www.qlik.com/

ていない．さらに、それらがデータの管理コストやセキュリティを考慮した上で適切に共有できる環境は確立されておらず，データに関する情報でさえ入手が困難な状況にある．

2.2 利用価値の高いデータはどちらか

一般的に、「無料のものよりも有料のものの方が質がいい」という先入観が存在する。データにおいても同様に、無料で入手可能なデータよりも有償のデータ（あるいはクローズドデータ）は質が保証されているという先入観があるかもしれない。それでは、実際にデータ市場において、一般に共有可能なデータと共有できないデータのどちらの方に利用価値あると認識されるのだろうか。

本論文では、データ市場を模したデータ利活用方法検討のためのワークショップ Innovators Marketplace on Data Jackets を用いることで利用価値の高いデータの特徴を抽出する。特にデータの共有条件に着目し、データ市場のステークホルダーのデータに対する潜在的なニーズについて考察を行う。また、データ利活用知識検索・推薦システム Data Jacket Store を用い、ユーザーに検索・閲覧されたデータの共有条件の抽出と比較を行う。

3. データ利活用方法検討ワークショップ：Innovators Marketplace on Data Jackets

データ市場創造を支援する手法として、大澤らは Innovators Marketplace on Data Jackets（IMDJ）を提案している[Ohsawa 13, Ohsawa 15]。IMDJ とは、データ市場に関わるステークホルダー間のコミュニケーションから、データ利活用案を検討する市場を模したワークショップである。

IMDJ では、データ自体ではなく、データについての情報（データジャケット）を共有することで、プライバシーの問題や情報漏洩のリスクを低減させている。データジャケット（DJ）は、データ自体は秘匿のままデータに関する情報、すなわちメタデータの一種である。この記述方法により、データの中身を公開することなくデータに関する情報を理解可能となる。さらに IMDJ では、DJ に含まれている情報から、他の DJ との潜在的なつながり可視化し、データの価値発見を支援する方法を導入している。例えば、テキストマイニングツール KeyGraph[Ohsawa 98]のアルゴリズムを用い、DJ として公開された「DJ 公開変数」から DJ 同士の関係性をシナリオマップとして可視化することができる（図 1）。IMDJ 参加者はシナリオマップから DJ 同士の関係を読み解き、データ利活用方法や組み合わせ案を考案する。

IMDJ では、参加者は要求を提示するデータ利用（消費）者、利活用案の提案者を兼任（DJ を提供している参加者はデータ保有者の役も担う）することで進行する。利用者の立場からは、自身が意識している問題を「要求」として提示し、提案者の立場からはシナリオマップ上の DJ を組み合わせることで利用者の要求を満たす「ソリューション」を創出する。これらのコミュニケーションにより、データ保有者は自身の保有するデータの利用方法を知ることができ、データを利用してビジネスを行いたいと考えている利用者、データ利活用方法を考案した提案者とともに、データの交換・売買、新ビジネス創造などの行動を開始する。

IMDJ 後は、創出されたソリューションを戦略的シナリオとして精緻化するアクション・プランニング[Hayashi 13]を実施する。要求分析及び実行動に関わる要素（ステークホルダー、リソース、データの変数など）の表出化と系列化により意思決定における盲点を低減させ、ソリューションの具体化を行う。

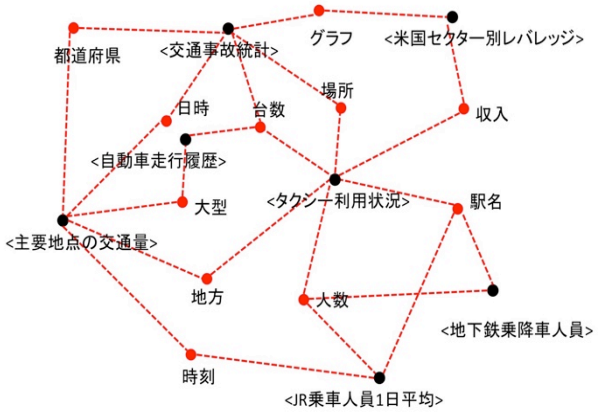

図 1 DJ 同士の関係性をシナリオマップとして可視化した例（黒ノードが各 DJ を表し，赤ノードに現れる DJ 公開変数を介し，DJ 同士がリンクでつながれる．）

4. Data Jacket Store（DJ Store）

オープンデータや Linked Open Data 関連研究では、セマンティック Web の基盤技術である RDF（Resource Description Framework）と SPARQL を用い、データに統一的にアクセスできる仕組みの構築が進んでいる。しかし、データの構造化と再利用は進んでいるものの、データ利活用の方法やそれによって満たされた要求についての情報は構造化され再利用されていない。

Data Jacket Store（DJ Store）は DJ を検索・推薦することでデータ利活用を促すための Web プラットフォームである[Hayashi 15, 早矢仕 15]。DJ Store には、どこにどのようなデータがあるのかという情報だけでなく、データ取得方法や分析方法、IMDJ において過去に創出された要求や検討されたソリューションも知識として構造化して RDF ストアに格納している。ユーザーは自然言語で文章やキーワードを入力し、関連するデータについての情報（DJ やデータ利活用案など）を取得することができる。DJ Store では、DJ だけでなく、過去に行われた IMDJ のデータ利活用方法と満たした要求も検索対象としている（図 2）。そのため、ユーザーは個人では気づき得なかった新しい問題解決方法やデータに関する情報を得ることができる。DJ Store は 2015 年 1 月から Web アプリケーションとして公開されており、現在も開発が進んでいる[DJS 15]。

また、DJ Store は分析ツールとしても機能する。例えば、過去の IMDJ でソリューション創出に用いられた DJ の一覧と利用回数を表示したい場合、クエリを DJ Store から RDF ストアに投げることで、欲しい情報の一覧を取得することができる。図 3 は SPARQL による問い合わせにより、DJ のランキングを作成した例である。

次章で説明する実験では、DJ Store の検索機能を DJ の特徴抽出に用いる。IMDJ においてソリューション創出に用いられた DJ に記述されている各データの共有条件（Sharing Policy）を DJ Store で取得し、比較を行う。

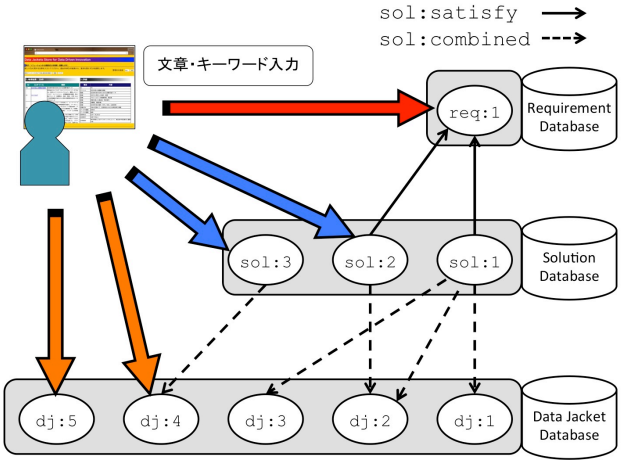

図 2　DJ Store から RDF で構造化されたデータベースにアクセスする。要求とソリューションは sol:satisfy という述語で連結されており、ソリューションと DJ は sol:combined という述語でつながっている。

図 3　DJ Store を用いて DJ のランキングを作成した例。Web ブラウザから SPARQL で RDF ストアにアクセスし（画面左）、一覧を取得する（画面中央）。

5. 実験

5.1 目的

データ市場において、オープンデータなどの一般に共有可能なデータと秘匿されたデータのどちらに価値があると認識されるかを DJ Store を用いて分析する。本実験では、IMDJ においてソリューションを創出するのに用いられた DJ を高評価 DJ と見なし、提案者のソリューション創出において有益と考えられるデータの特徴を調べる。特に、本実験では DJ の記述項目であるデータの共有条件に着目し、これを比較する。

また、補足実験として、IMDJ 中に DJ Store を利用したユーザーの検索行動から、ユーザーがソリューションを導くために閲覧した DJ の共有条件についても比較を行う。

5.2 方法と手続き

本実験において DJ Store に格納したデータジャケットは 426 件、知識として再利用する過去の IMDJ 情報として、335 件の要求と 298 件のソリューションを用いる（総トリプル数 11270 トリプル）。本実験の RDF ストアには、sparqlEPCU[*4] を用いた。また、DJ Store にユーザーが入力した文章をキーワードに分割する際に、「データ」や「情報」などの極端に頻度の高い名詞、「。」や「、」などの記号、助詞・助動詞は不用語として除外した。

2015 年 1 月 31 日から 3 月 13 日の 42 日間の DJ Store を Web 上に公開し、ユーザーのアクセスログを取得した。本実験ではこのアクセスログを用い、ユーザーの検索行動による閲覧 DJ と閲覧数を取得した。

5.3 比較方法

本実験で対象とした DJ Store 登録 DJ 数は 426 件、ソリューション 298 件に用いられた DJ 数は 186 件（重複を除く）、DJ Store の閲覧 DJ 数は 117 件（重複を除く）であった（表1）。そこで、発見件数の比較ではなく、出現割合を比較することとした。

表 1 DJ 発見数の比較

	件数
DJ Store に登録されている DJ	426
IMDJ でソリューション創出に用いられた DJ	186
DJ Store で閲覧された DJ	117

IMDJ でソリューション創出に用いられた DJ のうち、共有条件がspであるものの出現割合を$frequency(IMDJ, sp)$と表すとする。IMDJ でソリューション創出に用いられた DJ 集合をDJ_{IMDJ}、共有条件がspである DJ 集合をDJ_{sp}と表すと、IMDJ でソリューションに用いられた DJ のうち、共有条件が「一般に共有可（$sp = public$）」である集合は、$DJ_{IMDJ} \cap DJ_{public}$と表すことができる。よって、IMDJ でソリューションに用いられた DJ のうち、共有条件が「一般に共有可」であるものの出現割合$frequency(IMDJ, public)$は、式(1)のように表すことができる（用いる記号の意味を表 2 に示す）。

$$frequency(IMDJ, public) = \frac{|DJ_{IMDJ} \cap DJ_{public}|}{|DJ_{IMDJ}|} \quad (1)$$

表 2 記号の意味

記号	意味
DJ	DJ Store に登録されている DJ の集合
DJ_{IMDJ}	DJ Store に登録されている DJ のうち、IMDJ においてソリューション創出に用いられたものの集合
DJ_{DJS}	DJ Store に登録されている DJ のうち、DJ Store によってユーザーが閲覧した DJ の集合
DJ_{public}	共有条件（$Sharing\ Policy$）が「一般に共有可」である DJ の集合
$DJ_{private}$	共有条件（$Sharing\ Policy$）が「一般に共有不可」である DJ の集合
$\|X\|$	集合Xの元の個数

*4 http://lodcu.cs.chubu.ac.jp/SparqlEPCU/

本実験では、DJ Store に登録されている DJ の $frequency(DJ, sp)$と IMDJ でソリューション創出に用いられた DJ の$frequency(IMDJ, sp)$、DJ Store で閲覧された DJ の$frequency(DJS, sp)$を比較することで、提案者のソリューション創出において有益と考えられるデータの共有条件を調べる。

なお、共有条件における$public$とは、DJ の記述様式に含まれる「一般に共有可能」を意味し、$private$は「条件により共有可(必要に応じて交渉)」、「範囲を限定して共有可」、「購入により共有可」、「共有できない」を示す。

5.4 結果

登録 DJ、IMDJ にソリューション創出に用いられた DJ、DJ Store で閲覧された DJ の出現割合のそれぞれを「一般に共有可($public$)」と「一般に共有不可($private$)」に分けて表したのが図 4 である。

登録 DJ 件数における「一般に共有可($public$)」であるものの割合は約 3 割であるのに対し、IMDJ でソリューションに用いられた DJ 及び DJ Store で閲覧された DJ はおよそ 2 割と低い値を示した。一方で、登録 DJ 件数における「一般に共有不可($private$)」である DJ の出現割合は、IMDJ 及び DJ Store における出現割合の方が高い値を示していることが分かる。

以上の結果から、データ利活用方法を検討する IMDJ においては、一般に共有可能なデータよりも、共有が困難なデータの方がソリューションを創出するのに提案者にとって有益なデータである可能性が示唆される。

また、DJ 検索・推薦システム DJ Store においても IMDJ と同様に、登録 DJ と比較して一般に共有可能なデータの出現割合が低く、共有が困難なデータの出現割合が高い。つまり、一般に共有可能なデータよりも共有できないデータの閲覧数のほうが高いということが分かる。

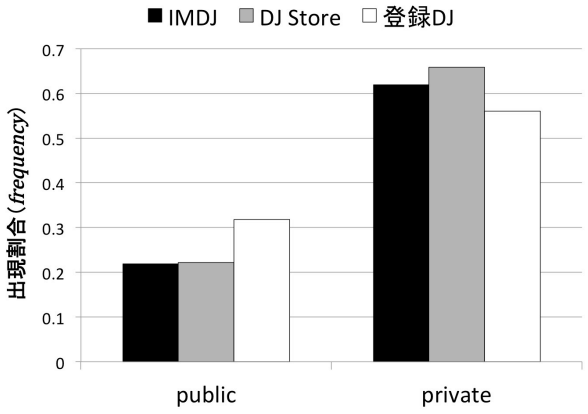

図 4 「一般に共有可($public$)」と「一般に共有不可($private$)」の出現割合比較

6. まとめ

本実験により、一般に共有が困難なデータほど、利活用方法を提案する上で有益なデータである可能性が高いということが分かった。つまり、オープン化できないデータほど、提案者及び利用者にとって問題解決及び新ビジネス創出において有用性が認められる可能性が高いということである。さらに、データに関する情報を検索するユーザーの検索行動においても、一般に共有不可能なデータの閲覧数の方が比較的高いという傾向から、共有が困難なデータの方がユーザーの興味・関心の度合いも高くなる可能性があることが分かった。

一方，高経年化原子力システムの安全性を議論する IMDJ において、データ利活用案(ソリューション)を見たデータ保有者は、該当データを共有するための有益な情報(所有者や入手方法など)を新たに提供する傾向が見られたという報告がある[大澤 14]. 以上の報告及び本実験の結果より、データ利用者及び提案者の一般にオープンにされてないデータ利活用への期待、そしてデータ保有者は自身の保有するデータ利用方法を知りたいというニーズが存在することが分かる。IMDJ ワークショップによるデータ利活用方法の提案により、データ保有者が自身のデータの利活用方法を認識すれば、積極的なデータの交換または売買が行われ、データ市場の活性化の可能性があると言えるだろう。

参考文献

[DJS 15] Data Jacket Store (Online): <http://www.panda.sys.t.u-tokyo.ac.jp/hayashi/djs/djs4ddi/ver.2/djs4ddi.html>, 最終アクセス 2015 年 3 月 25 日。

[Hayashi 13] Hayashi, T., Ohsawa, Y.: Processing Combinatorial Thinking: Innovators Marketplace as Role-based Game Plus Action Planning, International Journal of Knowledge and Systems Science, 4(3), pp.14-38, 2013.

[Hayashi 15] Hayashi, T., Ohsawa, Y.: Knowledge Structuring and Reuse System Design Using RDF for Creating a Market of Data, 2nd International Conference on Signal Processing and Integrated Networks, pp.566-571, 2015.

[早矢仕 15] 早矢仕晃章, 大澤幸生, データ利活用知識の構造化と再利用によるデータ情報推薦システム Data Jacket Store の提案, 電子情報通信学会 ライフインテリジェンスとオフィス情報システム研究会(LOIS), 信学技報, 114(500), pp.61-66, 2015.

[大向 13] 大向一輝: オープンデータと Linked Open Data, 情報処理, 54(12), pp.1204-1210, 2013.

[Ohsawa 98] Ohsawa, Y., Benson, N. E., and Yachida, M.: KeyGraph: Automatic indexing by co-occurrence graph based on building construction metaphor. In Proc. Advanced Digital Library Conference, pp.12-18, 1998.

[Ohsawa 13] Ohsawa, Y., Kido, H., Hayashi, T., Liu, C.: Data Jackets for Synthesizing Values in the Market of Data, 17th International Conference in Knowledge Based and Intelligent Information and Engineering Systems – KES 2013, Procedia Computer Science 22, pp.709-716, 2013.

[大澤 14] 大澤幸生: 市場メカニズムを模倣したシステム安全評価に資するデータ共有・活用手法の研究, 平成 25 年度高経年化技術評価高度化事業報告書, 2014.

[大澤 14] 大澤幸生: データジャケット ー創造的コミュニケーションのあるデータ市場のためにー, 人工知能, 29(6), pp.622-627, 2014.

[Ohsawa 15] Ohsawa, Y., Kido, H., Hayashi, T., Liu, C., and Komoda, K.: Innovators Marketplace on Data Jackets, for Valuating, Sharing, and Synthesizing Data, Knowledge-based Information Systems in Practice, Smart Innovation, Systems and Technologies, Springer-Verlag, Vol.30, pp. 83-97, 2015.