

論文の内容の要旨

論文題目 ユーザを介した複数公共データの統合による知識抽出に関する研究

氏 名 大知 正直

公共データの公開及び活用は、政府が2016年秋に設置した「未来投資会議」においても重要なテーマとしてあげられ、我が国の経済成長や社会課題の解決に貢献するものとして、重点的な推進が企図されている。公共データはこのように社会的にも大きな関心が寄せられている一方、現時点では限定的な利用にとどまっている。さまざまな公共機関、事業者が持つ公共データを、相互に提供しあうことによって、より大きな付加価値を引き出そうとしないのはなぜだろうか。本研究は、このような問いから始まっている。本論文では、公共データの活用が進まない現状に対する課題を、複数公共データの統合・活用に関する技術的な課題及びデータ保有者間の協力に関する障壁が存在していることにより、複数のデータを分析することで得られる知見の価値が、提供するリスクと比較し低くなっているものと捉え、技術的な課題の解決と障壁の低減により、リスクを上回る価値のある知見の抽出が可能となり、複数公共データの統合・活用が進むという仮説の検証を行っている。

そもそも複数のデータを持ち寄り、分析しようとする動機は何だろうか。それは各データを持ち寄り提供しようとする人々（公共機関、事業者）が分析をしようとする対象に何らかの共通項があると考えからであろう。本研究では、分析対象となるデータ群に共通する何らかの対象を単にユーザと呼ぶ。そして、このユーザに関する情報がどのように保存されているかは、データの保有者によって内容、形式ともに多様である。中には実際は、全くユーザが利用していない無関係なデータもある可能性がある。そこで、本論文ではユーザを中心としてデータを横断して分析しようとするのではなく、データソースを中心に分析を行う異種データ間のユーザ移動ネットワークという考え方を提案する。そして、データソース間を移動するユーザを介した複数公共データの統合による知識抽出を試み、検証を行う。もちろんデータソース間の保存内容、形式に差異が大きい場合、通常、知識抽出は困難になるため、データの形式を想定することで、それぞれの場合における知識抽出法の有効性に関する検証を行った。具体的には、データソー

空間でデータを生成したユーザを一意に特定できる場合と、一意には特定できないが部分的に特定可能な場合である。ここで部分的に特定可能であるとは、データ間を移動するユーザ群の持つ共通の属性、行動等で集団として特定可能な場合を指す。そして、それぞれの場合において、有効な知見を獲得するための分析の枠組みを示した。複数公共データの利用又はデータ提携に関しては、技術的な課題以外に、データ保有者間の協力に関して、情報の守秘、便益の不均衡といった課題が存在する。そこで最後に、実務的な仮説の検証を行うために、前章までで開拓した技法群の考え方を参考とした上で、それぞれの事業者の感じるリスクを解決しデータ提供を促すための調停機能を実装した独自のシステムを開発し、実際に各事業者の分析担当者の利用に供した上で評価を受けた。これらの検証を通じて、データを組み合わせることで得られる知見の価値がデータを他社に提供するリスクを上回れば、データによる事業者間提携が促進されることを明らかにした。

複数のデータソースから得られたデータを組み合わせるための統一的な枠組みは存在せず、また、実社会においてもデータを組み合わせる分析を行う取り組みはあまりされていない。一方で、複数公共データの統合による知識抽出は、事業者側の需要は大きいと考えられる。第1章では、多様なユーザが日々大量にデータを生み出しており、そのデータはプラットフォームを提供している各公的機関、事業者で分散されて保存されているが、この分散保存されているデータを集め、統合して分析するにはいくつかの課題があるために、有効に活用されていないことを説明する。そのような現状認識を踏まえ、本論文では、複数データによって得られる知見の価値が低く見積もられていること、データを提供することにリスクを感じていることの2点が課題であることを示す。本論文では、これらの課題の解決方法として、「ユーザを介した複数公共データの統合による知識抽出」手法を提案する。そのためにまず、獲得する知見の価値を高めるための方法の一つとして、異種データ間のユーザ移動ネットワークの提案を第3章で行う。この際、各公的機関、事業者間の持つデータの性質によって2つの場合に分ける。具体的には、データソース間でユーザを一意に特定できる場合と部分的に特定可能な場合とで、具体的な知識抽出を行うためのそれぞれのユーザ移動ネットワークの枠組みについて議論する。

第4章では、データソース間でユーザを一意に特定できないが部分的に特定可能な場合における知識抽出手法の提案とその有効性の検証を行った。ここでは、データ間で共通するユーザの興味、行動という点でユーザ移動ネットワークを構成する。このユーザ移動ネットワークを用いて、ヒット商品におけるユーザの検索行動が複数のメディアでの露出にどのように影響されているか、という分析を行っている。ここでは、人々の関心というものはメディアでの露出量と大きな相関があり、メディアで露出すると一定の割合でユーザの興味が喚起され検索行動を行うという過去の知見に基づいた仮定をおくことで、複数のデータソースをネットワーク化している。ソーシャルメディア上での

関心の盛り上がりや「口コミ指数」として指標化することを提案し、ソーシャルメディア内における関心の盛り上がりや、実際のヒット現象に与える影響を明らかにした。

次に第5章では、データソース間でユーザを一意に特定できる場合における分析手法の提案とその有効性の検証を行った。本章では、それぞれのユーザによるデータ間の移動の系列によって、ユーザ移動ネットワークを構成する。具体的には、公共交通の複数の運営会社から提供された乗降記録を用いて、ユーザの移動目的を推定する分析手法（「移動目的推定モデル」）を提案している。それにより特に、ユーザを特定できる移動ネットワークにおいては、文脈情報の利用を積極的にすることが移動目的の推定に重要な役割を果たしていることを示した。また、人々は、用事はなるべく近くで済まそうとするが、重要な目的がある場合には遠方へ移動する、等の知見を定量的な形で抽出することに成功した。

第6章では、事業者間のデータ提携における課題、利点を明らかにするため、前章までに示した分析手法の考え方を参考とし、独自の調停機能等を装備したシステムを開発し、その有効性と有用性に関して検証を行っている。実際に、複数の公共交通の運営会社からデータの提供を受け、それらを統合し、各社の担当者にとって興味を持つ分析を自由に行えるように設計を行った。その際に、それぞれの会社の持つ分析したい内容に対する希望と他社に開示したくない情報を把握した上で、便益と不利益の調整を試みている。実際の利用の後のアンケート調査を分析した結果、構築したシステムの有効性と、それぞれのデータ提供者が不利益にならないような設計（調停者の存在）が重要であることがわかった。

本論文では、複数公共データの統合・活用に関する技術的な課題及びデータ保有者間の協力に関する障壁を乗り越えるため、異種データ間ユーザ移動ネットワークという分析の枠組みを提案し、手法及び調停機能等を実装したシステムの実装とその有効性等に関する検証を行った。複数公共データの統合による知識抽出という複雑な課題に対し、一定の枠組みを示したことで、実際の事例において知識抽出の手法の開発、選定に注力することが可能になったと考えられる。今後、本論文の提案手法を用いることで、データを複合的に組み合わせることで得られる知見の価値が、自社のデータを他社に提供するリスクを上回るということについて社会で認知されるようになれば、データの事業者間提携がこれまでよりも促進されることが期待できる。