IoT eラーニング

ビッグデータの活用
（概要、動向）

国立大学法人　琉球大学

目次

ビッグデータの概要

● ビッグデータの定義

「分析を目的として大量の情報を収集し保管する」この手法自体は昔から行われてきたことである。「ビッグデータ」も「分析を目的として大量の情報を収集し保管する」ことと違いはないが、2000年代初頭に業界アナリストのダグ・レイニーがビッグデータの定義として３つのＶで表現した。

- 量（volume）
  第一の特徴として挙げられるのは、容量が大きいことである。情報技術の発展により様々なデータが大量に集まるようになり、データ量はテラバイトからペタバイトへと容量を増やしていっている。大量のデータを保管する技術も発展を続けており、その負担は軽減されてきている。

- 速度（velocity）
  製造工程のモニタリングシステムなどで使用されているセンサーやWebサーバのアクセスログなど多様性に富んだ大量のデータが、凄まじい頻度とスピードで生成され、取得と蓄積が行われている。この膨大な量のデータを効率よく分析するには、高速なデータ処理速度が重要となってくる。
  また、著しく状況が変わる現代社会ではリアルタイムに処理し対応することが求められている。ただし、蓄積されたデータは時間が経過することによって価値が下がるわけではないので、リアルタイム性は要素の一つであり全てではない。

ビッグデータの概要

- 多様性（variety）
  収集されるデータは、従来型のデータベースが扱うような「構造化データ」だけとは限らない。電子メールや音声、画像などの「非構造化データ」もあり、これらのデータを加工（テキストマイニングや構造化）し活用する動きが広まっている。

以上がビッグデータの「3V」として提唱されたものである。

ビッグデータの概要

「3V」に次の２つを加えて「5V」と提唱されることがある。

- 価値（value）
  ビッグデータは多様性に富む大量のデータと速度に価値があるわけではない。収集したデータを分析し有益な情報を掘り出し、モデルの構築・検証を繰り返し課題解決に導くことが本質的なビッグデータの価値となる。

- 正確性（veracity）
  従来は、全てのデータを集めることは出来ず、サンプリングによる一部のデータで全体を推測する方法を取っていた。ビッグデータでは全てのデータを取得することも不可能ではないため、推測による曖昧さや不正確などを排除した正確なデータによる意思決定が可能となる。

以上の特徴のように、どれだけのデータ規模なのかという量的側面だけでなく、どのようなデータで構成されているのか、またはデータをどのように利用するのかという質的側面において、従来の「分析を目的として大量の情報を収集し保管する」とは違うといえる。

ビッグデータの分析

- **分析目的の重要性**

「ビッグデータ分析でコストを削減したい又は顧客の数を増やしたい」等のしっかりとした分析目的を持たずにスタートしてしまうと失敗に終わってしまうことがある。

ビッグデータ分析の結果として「コスト削減」や「顧客の数を増やす」ということを実現することは出来る。
ただし、そのアプローチとして、
◆ コスト削減の対象はなにか、
◆ なぜコストが掛かっているのか
など、まず現状の課題を捉えたうえで、解決のための手段としてビッグデータ分析を用いる必要がある。

場合によっては、現状課題の解決にビッグデータという手段を必要としないかもしれない。

このため、
◆ 現状課題は何なのか
◆ なぜビッグデータ分析が必要なのか
などを考え、しっかりとした分析目的を持つ必要がある。

ビッグデータの分析

- **「相関関係」と「因果関係」**

データ分析を行う上で「相関関係」と「因果関係」を抑えておくことは必要なことである。

- 相関関係とは

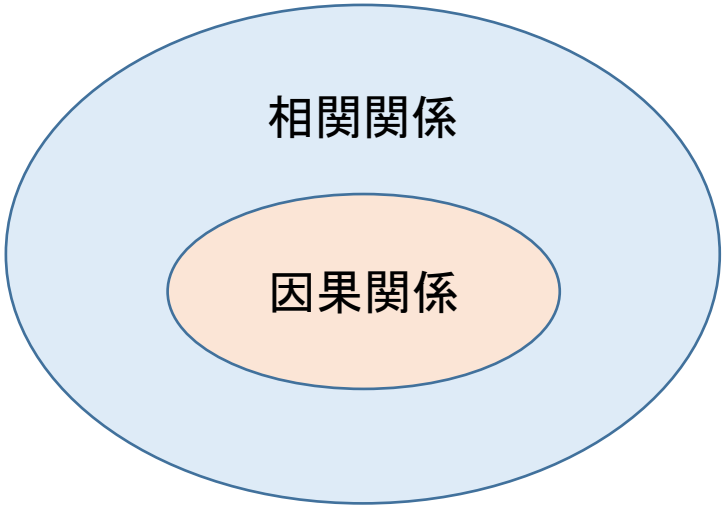
  相関関係とは、「一方の値が変化すれば他方の値も変化する」という２つの値の関連性を意味している。
  例えばX=2Yという数式の場合、Xの値が変わればYの値も変わると言え、Yの値が変わればXの値も変わると言える。
  また、時間の経過と太陽の位置には相関関係があると言える。
  朝の時間には太陽は東の空の低い位置にあり、太陽が東の空から西の空に移動すると朝の時間から夕方の時間まで時間が経過したといえる。
  相関には強さを持っており、「強い相関」や「弱い相関」と呼んでいる。例えの数式は逆も成立する強い相関と言える。

- 因果関係とは

  因果関係とは、ある2つ以上のものの間に原因と結果という関係があることである。
  例えば、「火災の規模が大きいと多数の消防士が出動する」の場合、「火災の規模が大きい」が原因となり「多数の消防士が出動する」が結果であると言える。
  この原因と結果の関係は一方向である。消防士の出動人数が増えることが火災の規模の大きさの原因とはならないからである。

ビッグデータの分析

「相関関係」と「因果関係」の関係を表すと次の図のようになる。

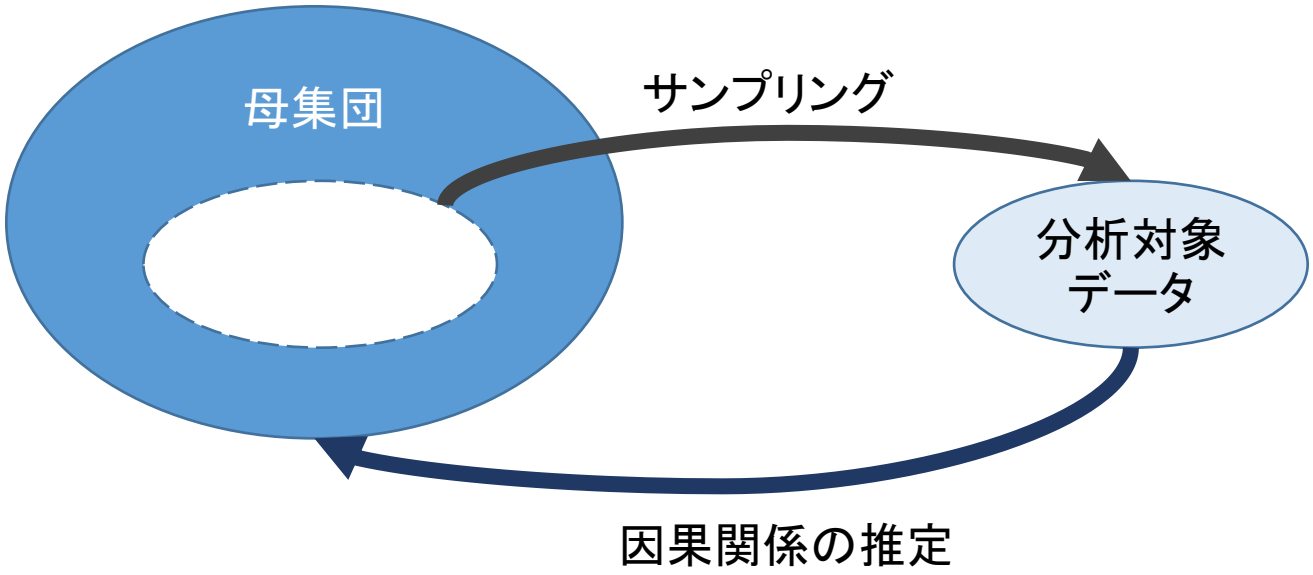


「相関関係」は「因果関係」を含んでおり、「相関関係」だけでは「因果関係」があるとは言えない。

ビッグデータの分析

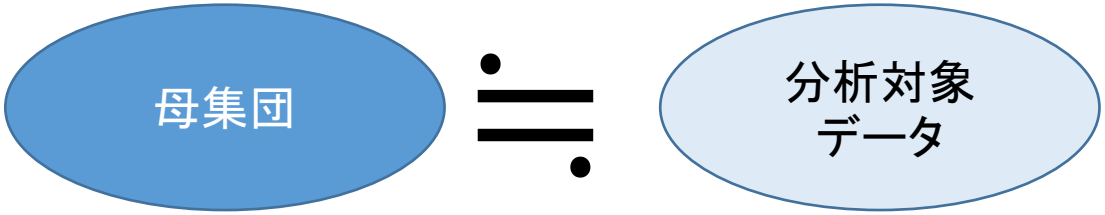
- **従来のデータ分析における「相関関係」**

従来のデータ分析では、分析対象となるデータは「母集団からサンプリングにより抜き出した一部のデータ」を使用している。このため、分析対象で求められたデータの関係は母集団でも適用可能なデータの関係であることを証明しなければならない。
これを満たすために、因果関係を見つけ出し妥当性をしめすことにより、母集団への推定へとつながることになる。
相関関係（強い相関）では、「ただの偶然」、「複数要因のうちの１つ」、「疑似相関」の場合があるため、母集団には適用できない可能性がある。



- **ビッグデータのデータ分析における「相関関係」**

ビッグデータの場合では、従来のデータ分析に比べて相関関係が非常に有用となる。
ビッグデータでは母集団とほぼ等しいデータを分析対象として扱えるため、「分析データで求められた関係」は「母集団に適用できる関係」であるといえる。因果関係を示さずとも母集団の関係を証明することができることとなる。
「分析対象データのすべての関係は、母集団にも適用される関係である」は重要なポイントとなる。



ビッグデータの動向

● ビッグデータの活用例

どのような分野でビッグデータを活用しているのか、いくつかの例を挙げる。

- 顧客動向の分析
  ECサイトにおいて、おすすめ商品を通知するサービスなどで活用されている。
  顧客が購買時に行った行動（だれが、何時、どこで、何を買った、等）をデータ化しビッグデータとして収集し、分析した結果をもとに個人毎へサービスを提供している。

- 多分野で活用される「GPS」
  スマホやカーナビで利用されている「GPS」データは、ビッグデータとして収集して活用されている。
  地図アプリの中で提供されているサービスとして、「GPS」の持つ位置と速度の情報をもとに渋滞情報を提供している。
  携帯会社ではスマホの「GPS」とアンテナとの「電波強度」などのデータを収集し、電波の弱い地域を把握、改善へと活用した。

ビッグデータの動向

- 医療分野での活用
  患者の治療内容や手術内容などの膨大なデータを持つ医療分野では、これらのデータをビッグデータ化し活用することが期待されており、がんに関するデータの公開などの活用が行われている。

- 金融分野での活用
  金融分野でもビッグデータの活用が進んでいる。
  取引情報をビッグデータとして収集、分析することで金融取引のリスクを回避したり、地域の公共機関などが提供する地域の経済動向や人口動態などと関連付け取引企業の経営をサポートしたりする等に利用されている。

ビッグデータの動向

- **ビッグデータのセキュリティ**

ビッグデータの中に蓄積されているデータには多種多様なデータが保管されている。

工場の工作機器に付けられたセンサーのデータや、スマホのビッグデータとして集められたデータであれば、電話番号、通話時間、通話内容、メールアドレス、訪問サイトなどが保管されていると思われる。

このように、ビッグデータとして取り扱われるデータは多種多様であり、その中にはある企業などにとっては重要なデータが含まれている場合もあれば、「個人情報」に絡むデータも多く含まれていることが考えられるため、データの取り扱いには十分注意する必要がある。

個人情報の代表として挙げられるのは「名前」「住所」「電話番号」といったものだが、場合によっては電話帳や表札などで調べることが可能なことから比較的秘匿性が低い情報ともいえる。

しかし、「嗜好」「思想」「宗教」などの情報は人によってはデリケートな意味を持ち厳重に秘匿すべき個人情報と言われている。

ビッグデータの動向

- **ビッグデータと個人情報**

JR東日本は2013年9月にSuica乗降履歴データの販売を当面見合わせると発表した。

これは同年6月27日に日立製作所がSuica乗降履歴を使った分析サービスを発表したことから端を発し、Suica利用者やマスコミなどから多数の問い合わせと国土交通省などを巻き込んだ社会問題となり中止へと追い込まれた。

この時、ユーザーへの事前通知がなかったことと、本人の申し立てで履歴の販売、譲渡を止められるオプトアウトの窓口を告知していなかったことが問題として騒がれたが、「個人情報」を匿名化し「匿名加工情報」へとする手法にも問題があった。

※「匿名加工情報」とは、「個人情報」を「匿名化」（加工）し元のデータには復元できないようにしたデータである。

ビッグデータの動向

JR東日本が販売しようとしたビッグデータは、Suica利用者の「ユーザID」を「ランダムなID」に変更しただけのものであった。さらに、置き換えた「ランダムなID」は、２年半の間「同一のID」でSuica利用履歴を追跡が可能となる杜撰なものであった。

「元データ」を保管するJR東日本を基準にした場合には「容易照合性」があり、このビッグデータは「匿名加工情報」ではなく「個人情報」である。

JR東日本はSuica利用者の「個人情報」を、告知することなく第三者（日立製作所）へ販売しようとしたといえる。

※「容易照合性」とは、他のデータと照合することで個人の特定ができる場合は「個人情報」であり、特定することができない場合は「匿名加工情報」ということである。

ビッグデータの動向

- **不十分な匿名化**

Suica乗降履歴データの件以外にも、不十分な匿名化によって問題が発生した事例がある。

インターネット事業者が検索ワードの一部をインターネット上に公開した際に、検索クエリに付属している「ユーザーID」を置き換えて「匿名化」の加工を施していた。
しかし、検索ワードに含まれている「個人名」「住所」「社会保障番号」などは何も加工せず、容易に個人を特定可能なデータを含んでいた。

また、アメリカの動画配信事業者が自社のユーザーが記入したレビューデータを匿名化して公開した。
しかし、ある大学の研究グループがインターネット・ムービー・データベース（IMDb）のレビューと公開されたデータを比較し、一部の個人を特定することに成功してしまった。
研究グループはレビューには「くせ」があると考え比較した結果、特定に至っている。

- **ビッグデータ活用時の個人情報保護**

ビッグデータには個人情報が大量に含まれていることがあり、その取り扱いには十分に注意を払う必要がある。

「個人情報」を「匿名加工情報」へと加工することの難しさは先の説明からも明らかである。

また、国毎に個人情報保護制度が策定されており、取り扱うデータと提供するサービスについては十分に配慮する必要がある。