情報の倫理

2017/12/07

Kazuma Sekiguchi
class@cieds.jp

Q1

- 飛行機が2倍以上も遅れる原因として多いのは次のうちのどちらか？

  - 雪
  - 霧

正解は

ビックデータとは

- 「事業に役立つ知見を導出するためのデータ」
  （情報通信白書2012年度版）

- 実際のところ明確な定義は無い
- 「ビッグデータは、高ボリューム、高速度、高バラエティの情報資産のいずれか（あるいは全て）であり、新しい形の処理を必要とし、意思決定の高度化、見識の発見、プロセスの最適化に寄与する」（ガートナーの定義）

ビックデータとは

- 雑多なデータからデータマイニングなどの統計的手法を用いて、これまでに無い新しい知見または将来的な予測をするためのデータ及びその活用法
  - 辺りが定義になる

ビッグデータ

- 旧来では無視されてきたデータを大量に蓄積し、分析することで新たな知見を得ようとする
  - 大量のデータ＝ビッグデータ
- 旧来はコンピュータの処理に時間が掛かる
  - データの収集にも多数の時間、コストが掛かった
  - データ量を抑え、少ない確実なデータで傾向を見つける
- 現在はコンピュータの処理が高速化
  - データの保存コストもほぼ無視できるレベルまで最小化されている
  - 使える全てのデータを分析し、何らかの相関関係を見つける
  - データに多少の誤差があっても気にしない（数の多さでその誤差をカバーすることが可能）

旧来のサンプル収集

- アンケート調査、訪問調査、電話調査、郵送調査
  - コストも時間も掛かりすぎる
- 標本の抽出も大変
  - 標本＝ある調査対象の母集団から、無作為に抽出した小数の対象
    - 単に分析のコスト、時間の問題から小数を抜き出して、全体を知ろうとした
    - 無作為抽出が確実にできれば、そこから得られたデータは十分信用に足る
  - 実際のところ無作為抽出が難しい
    - 職業、年齢、性別、住所、年収などさまざまな条件を勘案する必要

データ処理量

- Google：1日に24PB（ペタバイト）
  - 24PB＝24,000,000TB=300億個の8GBUSBメモリ
- FaceBook：「いいね」とコメントは1日30億回
- YouTube：月間利用人数が8億人
  - 1秒に付き長さ1時間分の動画を投稿している
- Twitter：1日に4億ツイート

現在のデータ収集

- ウェブのクリック数、滞在時間、どこのウェブサイトを見ていたか、使ってるブラウザーの種類、アクセスしてきている大体の場所、スマートフォンの機種名などなど
- 検索時にどういう言葉を入れてどのサイトを実際に見ているか
- ツイッター、Facebookの情報

現在のデータ収集

- さまざまなセンサー
  - 温度、振動、回転数など
- GPS
  - 位置情報
- Suicaなどの交通系ICカード
- Tポイントカードなどのポイントカードシステムの利用
- 携帯電話の電波状況

現在のデータ収集

- 監視カメラのデータ
  - 人物をあるレベルで識別可能
- ウェラブルデバイスのデータ
  - 身につけてデータを取得するためのデバイス
  - 健康目的などで活用される

- ありとあらゆる情報を集めて処理するだけの能力を有する

ビックデータにおけるデータの特性

- 3V
  - Volume(データ量)
  - Variety(種類)
  - Velocity（リアルタイム性）

の3つがビックデータでのデータとして有益なもの
（ガートナー）

- 粒度による切り口も考えられる
  - 個人の詳しい情報を集積
  - 匿名の大量のデータを集積

ガストでのメニュー開発

- 「若者はハンバーグ、シニア層は和食」という推定
  - シニア層もハンバーグを注文している
  - 「チーズＩＮハンバーグ」は他の商品よりもリピート率が高く、幅広い層が支持する看板商品

- 計算に基づくメニュー提供
  - 中年男性には鉄板に乗ったステーキ
  - 中年女性には白い皿に盛ったピラフとステーキのメニュー

- 勘に基づかないメニュー開発

スシローでの寿司提供タイミング

- お客の年齢、組み合わせにより、お客が要求する可能性の高い時間にあわせて適切な寿司をレーンに流す
  - 「着席して１分後に何を食べそうか」「そのあと１５分後まで何を食べるか」という予測が瞬時に厨房に表示
- 廃棄率が1/4以下に減少

- 皿の裏側に付けられているICタグを利用し、どのネタが誰に何分後に取られたかを把握していく
  - 大量のデータを集計して、予測に活かす

インフルエンザ流行の予測

- Googleがほぼインフルエンザの流行と同時にどこで流行し始めているかの場所を州単位で特定した

- 上位5000万件の検索語を抽出し、インフルエンザ流行のデータと照合
  - 4億5000万の数式モデルを組み合わせて、相関関係の有意を調べた
  - 特定の45検索語とある特定の数式モデルの組み合わせで、インフルエンザ流行との高い相関関係を発見
- これをそのまま適用することで、ほぼリアルタイムにインフルエンザの流行を州単位で特定

コミュニティ内の交流

- 携帯電話の通話4ヶ月分を全て分析
  - 小規模のサンプル調査と異なる結果
  - コミュニティ内で多くの接点を持つ人（ハブ的存在）がいなくなっても交流程度は低下するものの交流自体は継続
  - コミュニティ外部に接点を持つ人がいなくなると残った人はコミュニティが崩壊したように交流が途絶える

  - 集団内の交友関係を盛り上げているのはハブ的存在では無く、集団外部と繋がりを持つ人間が盛り上げ役
    - これまでの調査と異なる結果

機械翻訳

- 日本語を英語に、またはその逆を自動的に
  - 現時点でも綺麗な日本語訳や英語訳は出てこない
- 旧来のアプローチ
  - 辞書による単純な翻訳
    - 「Hello」→そのまま辞書的なら「こんにちは」
      - 現実的には、「やぁ」かも、「もしもし」かもしれない
    - 言葉には例外が多い
  - 言葉のルールを覚えさせる
    - 300万語センテンスを追加して翻訳させる
    - 成果は今ひとつ

機械翻訳

- Googleによるアプローチ
  - きちんと翻訳された2カ国間の資料だけでは無く、ネット上のありとあらゆるデータを利用
    - 企業のウェブサイト、公文書、書籍の翻訳など
    - 950億センテンスを利用
  - データ量を増やし、確率を判断に翻訳を実行
    - 比較的精度の高い翻訳が可能になった
  - ネットのデータを利用するため、最近の言葉でも翻訳に対応出来る利点

Amazonのお勧め商品

- Amazonは創業以来多種のデータを保持
  - 最後まで迷ったが購入に至らなかった書籍、どの本をどのくらいチェックしていたか、一緒に購入した書籍はどれであるか、など
  - 旧来はデータを利用せずに関連商品を勧めていた
    - フランスに関する本を買ったら、「フランスガイドブック」やら「フランス料理」などを勧める
  - 前回と大差ない商品を永遠と勧めるはた迷惑な店員

- 関係なさそうな商品同士の相関関係に基づく推薦を実装
  - 100倍以上商品が売れるようになった
  - 現在は1/3がお勧め商品とパーソナル化システム

人の感情を理解するロボット

- 「Pepper」
  - すべての感情を理解させることは現状では不可能
    - 昔はすべての感情を理解させようとして失敗した
  - 表情や声などから感情を数値化して学習
    - 良い感情表現や悪い感情表現を溜めていく
    - クラウド上でPepper同士の情報を集積、分析をすることで、より正しい方向へ向かわせることが可能になる
- 大量のデータを個々に学習させ、数値化（標準化）した上で集積、個々へ学習データとして戻すことによる精度向上を狙う

Q2

- アメリカのハリケーンが近づくと懐中電灯とともに売れる商品は何か？

  - 傘
  - お菓子
  - かなづち

正解は

お菓子

ポップターツという甘い
お菓子が売れる

因果関係ではなく相関関係

- 人間は世の中を因果関係で見ている
  - 手っ取り早く架空の因果関係を持ち出す
    - 「世の中の景気が悪い」→「物が売れないから」→「若者が物をほしがらなくなった」
  - 本当の原因が分からなくても、「因果関係を知りたい」という本能的な欲望に基づく
    - 実際のところは分からない
    - 「レストランに行った」→「お腹が痛くなった」→「あのレストランで食べたものが悪かったに違いない」

Q3

- 中古車として購入した場合、もっとも故障が起きる可能性が低い車の色は何色？
  - 青
  - オレンジ
  - シルバー
  - 赤
  - 白

正解は

オレンジ

因果関係を考えない

- 「中古車オークション」に出品されている車の品質に問題がありそうな車を予測するアルゴリズム
  - アルゴリズム＝ある問題を解くための手順を定式化したもの
  - 中古車ディーラーから提供されたさまざまなデータを元に分析を行いアルゴリズムを考案

- 全てのデータから相関分析を実施したところ
  「オレンジ色の車は欠陥が大幅に少ない」

- 因果関係を考えても無駄な例
  - 相関関係としては出てくるが、因果関係としては説明が付かない

因果関係を考えない

- 自動販売機での飲み物の販売
- 人が目に付く場所は自動販売機の左上、というのが常識

- 自動販売機にアイトラッキング（視線がどこを向いているか調べる装置）を取り付けてデータ採取
- 下段の左端を見る傾向が強い

データの収集

- スマートフォンアプリなどでも使用しないにもかかわらず位置情報を取得するものが多数
  - 位置情報を利用して、新しいサービスを展開する
- 携帯電話会社は電波受信状態を常に監視して改善策に活かすことをしている
- 居場所が分かれば、ピンポイントで広告を打つことも可能
  - トレンドを掴むことも可能

データの収集

- Facebook
  - 多数のデータおよび人間関係までもデータとして保持
  - 10億人が利用するメディア
  - 現時点ではデータの利用は明確に言われていない
- Twitter
  - 自社でデータの活用はしていない
  - ツイートされたデータを他社に提供
    - 多くの企業で感情分析に利用（喜怒哀楽を文字情報から読み取る）
  - Twitterはツイート以外にユーザ情報に関する33項目を提供している
    - 言語、利用位置、フォロー先、フォロワー数など

データの収集

- 交通系ICカード
  - Suicaなど
- ショップのポイントカード
  - Tポイント、nanaco、Ponta辺りが熱心とされる

- 監視カメラのデータ
  - 人物をあるレベルで識別可能
- ウェラブルデバイスのデータ
  - 身につけてデータを取得するためのデバイス
  - 健康目的などで活用される

- ありとあらゆる情報を集めて処理するだけの能力を有する

データ収集

- Amazon
  - 購入した書籍、単に眺めたページなどを記録
- Google
  - 検索語、検索時に利用した位置、PCの環境、何番目をクリックしたか、何秒ページを見たかなど
  - GoogleAnalyticsという分析ツールの提供
    - Webページの管理者がアクセス数などを把握できるツール
    - Googleは当然そのページに何人の人が来たかなど、ツールを利用しているウェブページから取得可能
  - Gmail
    - メールの内容、友達

「もしかして・・・」

- Goolgeでの検索時にスペルミスしたり、検索語が間違えていたり、件数が少なかったりした場合に表示される
  - 最近では、検索語ではなく、修正後の言葉で検索されることも発生
  - 「もしかして」に付随して、キーワードが付いていることも
  - そのままユーザがページにアクセスすれば、新しい言葉やスペルミスを更に修正できる

  - Googleサジェスト機能訴訟
    - 自分の名前を検索すると犯罪情報が出てくることから不利益を被る