インターネット計測とデータ解析 第 14 回

長 健二朗

2014 年 7 月 14 日

前回のおさらい

第 13 回 スケールする計測と解析 (7/7)

- 大規模計測
- クラウド技術
- MapReduce
- 演習: MapReduce

今日のテーマ

第 14 回 まとめ

- インターネット計測とプライバシー
- これまでのまとめ

プライバシー

他人の干渉を許さない、各個人の私生活上の自由 – 広辞苑

みだりに自分の私生活を公開されない権利、法的保証
個人の情報を自分でコントロールできる権利

- プライバシーの見方はコンテキストや文化で大きく異なる
  - 基本的人権
  - 財産権: 個人情報が商品価値を持つ。侵害されれば損害賠償

プライバシー情報

- サービス利用履歴、web 閲覧検索履歴、購入商品、趣味指向
- 本人が自ら公開している場合はプライバシー情報とはならない
- しかし、情報の収集、加工、第三者への提供などもプライバシーの侵害になりえる

個人情報

個人を識別することができる情報

- 氏名、性別、生年月日、住所、電話番号、家族構成、職業、年収、生体情報
- IP アドレス、メールアドレス、オンライン上の ID、位置情報
- 財産情報、購買履歴、医療記録、趣味嗜好、宗教情報、写真、通信記録
- 日本の個人情報保護法　2005 年に施行
  - 5000 件以上の個人情報を扱う事業者が対象
  - 利用目的の特定、制限、適切な取得、通知義務、苦情処理

個人情報を含む記録

- キャッシュカード、クレジットカード、交通系 IC カード、メンバーカード
- デバイス ID: SIM カード、MAC アドレス、IP アドレス、RFID
- Web クッキー、位置情報、 監視カメラ、指紋、顔認識

総務省パーソナルデータに関する検討会 (2013-2014)

- パーソナルデータの利活用ルールの明確化と制度の見直し
- データの利活用戦略とプライバシー保護や国際連係とのバランスを検討

誰が個人情報を持っているか

公的機関

- 政府系組織
- 病院
- 銀行
- 大学

商用サービス

- 店舗、その他のサービス
- ソーシャルネットワークサービス

市場価値

- 人口統計データ、位置情報、その他の統計 (個人が特定できない条件)
- ブラックマーケットの存在 (データ盗難などの犯罪)

インターネット計測とプライバシー

計測はすべての技術の基本
計測情報の開示: 個人情報を含まない統計情報のみ開示可能
計測データからプライバシー情報が漏洩するリスク

- 計測データ中の個人情報 (IP アドレスなど)
- 技術の進歩で情報の拡散や加工が容易になった
- 悪意の利用やリバースエンジニアリングの可能性

技術に法制度がついていけない現状

- ほとんどがインターネット以前に作られた制度
- 計測には法的にはグレーな部分が多い
  - 計測に対する立場の違い、技術者の認識にも大きな温度差

通信の秘密

憲法上の通信の秘密

- 政府など公権力に対する義務

電気通信事業法第 4 条第 1 項で通信の秘密

- 電気通信事業者の取扱中に係る通信の秘密は、侵してはならない

例外

- 当事者の同意がある場合
  - ウイルスチェックサービスや迷惑メールフィルタリングサービス
- 違法性阻却事由が存在し、違法とはされない場合
  - 業務上必要な正当業務行為に当たる場合
    - 例: パケット配送のためにヘッダ情報を見る
  - 緊急避難に該当する場合
    - 例: 他のサービスに支障が出ないよう対策をする

インターネット計測とプライバシー漏洩リスク

生データ、汎用データ

- 当初の目的以外の利用が可能、情報漏洩リスクを伴う
- 汎用性と情報漏洩リスクのトレードオフ
  - 例えば、特定目的用にオンライン処理することでリスク低減

データの共有、公開

- 共有: 第三者への情報提供となる問題
  - 必要最小限の情報のみ共有するようなデータの加工は可能
- 公開: 幅広い利用促進、悪用されるリスク

商用トラフィックと非商用トラフィック

- 研究教育用ネットワークは比較的計測しやすい
- いっぽうで、商用トラフィックとの乖離

インフォームド コンセント

- 利用者に説明、理解と合意を得るプロセス
- 医療分野で進んでいる (倫理委員会設置など)

法的側面とモラル

- 合法であるかだけでなく、技術者のモラルが問われる
  - センシティブなデータの削除や匿名化

例: Netflix Prize

- 米国のオンライン DVD レンタルサービス Netflix のアルゴリズムコンテスト
- 同社のオンラインお薦めシステムの性能を 10%向上すれば 100 万ドルの賞金
- コンテスト用データセット:
  $< user\_id, movie\_id, date\_of\_grade, grade >$
  - トレーニング用データセット: 1 億件の評価スコア
  - 評価用データセット: 280 万件の評価スコア
    - 答え合わせ用データセット: 140 万件
    - 採点用データセット: 140 万件
  - 採点スコアは結果の誤差の平均二乗偏差 (10%改善目標)
- コンテストは 2006 年に始まり、2009 年に終了
  - プライバシー問題で批判
  - 匿名化されたユーザを他の映画評価サイトのユーザとマッチング可能

匿名化した情報とプライバシー漏洩リスク

行動履歴

- 位置情報、移動履歴
- 購買、閲覧履歴
- マーケティング応用の価値
- 以前は、匿名化すれば個人情報ではなく第三者への提供も可能
- 情報技術の進歩で、個人特定のリスクが増大

個人を特定できる情報

- 名前、会員番号、電話番号など直接個人と紐づけされる情報

個人を識別できる情報

- 匿名化された誰かひとりの情報
- 他の情報と突き合わせて個人を特定できる可能性
  - 時間、場所、希少レコードの照合

- 例: 大学キャンパスでの匿名アンケート
  - 慶応義塾大学 学生数 33,681 (男 22,499 女 11,182)
  - SFC 学生数 4,851 (男 2,871 女 1,980)

安全・安心とプライバシー

- デジタル著作権管理 (DRM: Digital Rights Management)
  - デジタルデータの著作権保護技術の総称
  - 著作権者、ライセンス管理者、消費者の立場
- テロや犯罪との戦い

- 利便性とプライバシー
  - サービスの統合
  - ソーシャルネットワークで友人からプライバシーが漏洩するリスク

プライバシーの今後

- ポストプライバシーの時代?
  - 将来はプライバシーの概念が変わってしまう可能性
  - プライバシーの歴史は意外に短い
    - 1890 年頃にマスメディアが登場した後
- 問題は複雑 (文化的、法的、経済的側面)
- ユーザは、自分で自分のプライバシーを守る必要
  - 悲観的になる必要はないが
  - 理解と認識が大切

授業のまとめ

授業のねらい
(学生に身につけて欲しいこと)

- データのばらつきについて理解し、データ処理とグラフ化を習得
  - 卒論や他のレポートを書くときに役立つはず
- 大量データを処理するプログラミング技術を習得
  - 既成のパッケージソフトウェア依存では限界
- 統計データを疑う力をつける
  - 作為的な統計データや情報操作の氾濫
  - (オンラインプライバシーに関するリテラシー向上)
- データ処理の肌感覚
  - プログラミングやデータ処理を通して初めて身に付く感覚

科目概要

インターネットによって、多様で膨大なデータが容易に取得できるようになった。そこから知見を引出し、新たなサービスを作り出すことが可能になり、ビッグデータや集合知として注目されている。しかし、これらを正しく理解し、道具として使いこなすためには、その背景にある統計、機械学習、システムに関する総合的な理解が欠かせない。

本授業は、インターネット上でのデータ取得と大規模データ解析の概要について学び、情報社会で必須となる大量情報から新たな知識獲得をするための基礎能力を身につける。

主題と目的／授業の手法など

インターネット上でのデータ収集とその解析手法について学習し、ネットワーク技術と大規模データ処理の総合的な知識と理解を得る。授業では、具体的な応用例について、その基礎技術と背景にある理論を関連づけて理解する。講義に加えて、毎回データ処理の演習を行い、習った理論をプログラムに実装してデータ処理をすることで、データ解析手法を身につける。

big data

- big data: 大量の非定型データから隠れた価値のある情報を引き出す技術の総称
  - 新たなビジネスモデルの構築や経営改革に繋げる
- big data という言葉をいたるところで聞くようになった
- 技術は以前から使われている
  - 検索ランキング、オンラインストアのお勧めシステムなど
  - インターネット計測: 大量かつ不完全なデータからインターネットを把握する試み
    - 統計的な手法による推測
    - 工学的な計測との対比

クラウドサービスの登場

- 以前は大量のデータの利用は、インハウスで収集、管理、分析ができる組織に限られていた
- クラウドサービスの普及で、誰でも大量データが使える環境が出来てきた
- 顧客のオンライン行動履歴を収集分析するパッケージツールも利用可能に
- 僅かな初期投資で顧客情報をマーケティング利用可能になった

データの時代

- あらゆる分野でデータ革命と呼べる技術革新が進行中
  - それまで難しかった応用が可能に
    - 膨大なデータへのアクセス、常に更新されるデータを対象にした解析、非線形モデルへの応用など
- あらゆる科学技術分野で、膨大なデータ解析は欠かせない研究手法になった

データ分析はあくまで道具

- 最近のビッグデータの話題はツールや手法が強調されがち
- データ解析はあくまでツール
  - 仮説を立てて、データで検証
  - 結果が予想と異なれば、そこから新たな疑問へ
  - このプロセスの繰返しから、役立つ情報や興味深い事実の発見
- 目的を持たずにデータを集め CPU を回し解析してもムダ
- 逆にデータから何を得たいかがはっきりすれば、やるべきことは見えてくる

思考プロセスの変化

- もちろん以前からデータを基に考えることは重要だった
- 情報技術によって、データに基づいて考え、考えをデータで検証する思考プロセスに変化
  - 扱えるデータの量と質、その表現方法が桁違いに
  - 文字通りデータと対話しながら考えることが可能に

データ時代の課題

- 人材 (データサイエンティスト) の育成
  - その分野の専門知識を持った上で、既存の考えや解釈に疑問を持つ、統計やデータ解析を道具として使いこなして問題解決をする
- データの財産化
  - 他社が持っていないような実データを持つ会社が強い
  - 同じデータなら、情報を引き出す能力で優劣
- データの共有
  - データを共有できる、検証できることの社会的意義
- プライバシーとのバランス: 社会的合意形成が大きな課題
  - 組織がどこまで個人を追跡していいか
  - 個人の医療情報などをどのように共有して社会に役立てるか

受け取りてのリテラシ

- 受け取り側も、データを理解する、データに疑問を持つ必要
  - 発信者のバイアスによる作為的な統計データや情報操作の氾濫
- 我々は白黒の判定を求めがち
  - ほとんどの物事はグレー、白黒は便宜的にグレーに線を引く行為
  - 白黒を求めるのは、自ら判断することを避けて、発信者に判断の責任を求める行為
  - グレーはグレーとして受け取り、自分で判断することが必要な社会になってきている

前回の演習: WordCount in Ruby

Ruby で MapReduce ぽい処理をしてみる

```
% cat wc-data.txt
Hello World Bye World
Hello Hadoop Goodbye Hadoop
% cat wc-data.txt | ruby wc-map.rb | sort | ruby wc-reduce.rb
bye     1
goodbye 1
hadoop  2
hello   2
world   2
```

WordCount in Ruby: Map

```
#!/usr/bin/env ruby
#
# word-count map task: input <text>, output a list of <word, 1>

ARGF.each_line do |line|
  words = line.split(/\W+/)
  words.each do |word|
    if word.length < 20 && word.length > 2
      printf "%s\t1\n", word.downcase
    end
  end
end
```

WordCount in Ruby: Reduce

```
#!/usr/bin/env ruby
#
# word-count reduce task: input a list of <word, count>, output <word, count>
# assuming the input is sorted by key
current_word = nil
current_count = 0
word = nil

ARGF.each_line do |line|
  word, count = line.split

  if current_word == word
    current_count += count.to_i
  else
    if current_word != nil
      printf "%s\t%d\n", current_word, current_count
    end
    current_word = word
    current_count = count.to_i
  end
end
if current_word == word
  printf "%s\t%d\n", current_word, current_count
end
```

最終レポートについて

- A, B からひとつ選択
  - A. ウィキペディア日本語版の Pageview ランキング
  - B. 自由課題
- 8 ページ以内
- pdf ファイルで提出
- 提出〆切: 2014 年 7 月 28 日 (月) 23:59

最終レポート 選択テーマ

A. ウィキペディア日本語版の Pageview ランキング

- ねらい: 実データから人気キーワードを抽出し時間変化を観測
- データ: ウィキペディア日本語版の Pageview データ
- 提出項目
  - A-1 Pageview カウント分布調査
    - 各ページの 1 週間分のリクエスト総数を集計し、分布を CCDF でプロット
  - A-2 各日および 1 週間合計からリクエスト数トップ 10 を抽出
    - トップ 10 の結果を表にする
  - A-3 週間トップ 10 についてランキングの推移をプロット
    - ランキング変化が分かり易いよう時間粒度を考え図を工夫する
  - A-4 オプション解析: その他の自由解析
  - A-5 考察: データから読みとれることを考察

B. 自由課題

- 授業内容と関連するテーマを自分で選んでレポート
- 必ずしもネットワーク計測でなくてもよいが、何らかのデータ解析を行い、考察すること

最終レポートは考察を重視する

課題 A. ウィキペディア日本語版の Pageview ランキング

データ: ウィキペディア日本語版のデータ Pageview データ

- wikimedia が提供するデータからウィキペディア日本語版だけを抜き出したもの。
- 元データ情報: http://dumps.wikimedia.org/other/pagecounts-raw/
- 課題用 Pageview データ: 20140616-22.zip (609MB 解凍後 3GB)
  - 1 時間毎の Pageview データ 1 週間分 (2014 年 6 月 16 日-22 日)
- オプションデータセット: 20140601-15.zip (1.3GB 解凍後 6.3GB)
  - オプション解析で利用可能な追加データ (2014 年 6 月 1 日-15 日)

データフォーマット

- project encoded_pagetitle requests size
  - project: wikimedia のプロジェクト名 (課題用データでは全て"ja")
  - encoded_pagetitle: URI エンコードされたページタイトル
  - requests: ページのリクエスト回数
  - size: ページのバイト数

```
% head -n 10 20140616-22/pagecounts-20140616-00
ja $ 1 0
ja $10 1 8922
ja %22B%22ORDERLESS 1 13777
ja %22BLUE%22_A_TRIBUTE_TO_YUTAKA_OZAKI 1 21159
ja %22HAPPY%22_Coming_Century,_20th_Century_Forever 1 21326
ja %22LUCKY%22_20th_Century,_Coming_Century_to_be_continued... 1 0
ja %22X%22_plosion_GUNDAM_SEED 2 50386
ja %26C 1 16485
ja %26_(%E4%B8%80%E9%9D%92%E7%AA%88%E3%81%AE%E3%82%A2%E3%83%AB%E3%83%90%E3%83%A0) 1 12635
ja %26_(%E6%BC%AB%E7%94%BB) 1 0

% head -n 10 20140616-22/pagecounts-20140616-00 | ./urldecode.rb
ja "$" 1 0
ja "$10" 1 8922
ja ""B"ORDERLESS" 1 13777
ja ""BLUE"_A_TRIBUTE_TO_YUTAKA_OZAKI" 1 21159
ja ""HAPPY"_Coming_Century,_20th_Century_Forever" 1 21326
ja ""LUCKY"_20th_Century,_Coming_Century_to_be_continued..." 1 0
ja ""X"_plosion_GUNDAM_SEED" 2 50386
ja "&C" 1 16485
ja "&_(一青窈のアルバム)" 1 12635
ja "&_(漫画)" 1 0
```

タイトルのデコードスクリプト

- タイトルはパーセントエンコードされている
  - ruby の CGI.unescape() で UTF-8 に変換できる

```
#!/usr/bin/env ruby

require 'cgi'

re = /^([\w\.]+)\s+(\S+)\s+(\d+)\s+(\d+)/

ARGF.each_line do |line|
  if re.match(line)
    project, title, requests, bytes = $~.captures
    decoded_title = CGI.unescape(title)
    print "#{project} \"#{decoded_title}\" #{requests} #{bytes}\n"
  end
end
```

課題 A Pageview ランキング補足

- A-1 Pageview カウント分布調査
  - 各ページの 1 週間分のリクエスト総数を集計し、分布を CCDF でプロット
  - X 軸はリクエスト数、Y 軸は CCDF、log-log でプロット
- A-2 各日および 1 週間合計からリクエスト数トップ 10 を抽出
  - トップ 10 の結果を表にする

```
rank  6/16 6/17 6/18 6/19 6/20 6/21 6/22 total
  1   "a"  "b"  "c"  "d"  "e"  "f"  "g"  "x"
  2   "h"  "i"  "j"  "k"  "l"  "m"  "n"  "y"
...
 10   "o"  "p"  "q"  "r"  "s"  "t"  "u"  "z"
```

- A-3 週間トップ 10 についてランキングの推移をプロット
  - X 軸に時間、Y 軸にランキングをとる
  - ランキング変化が分かり易いよう時間粒度やランキングの見せ方を考え図を工夫する

まとめ

第 14 回 まとめ

- インターネット計測とプライバシー
- これまでのまとめ