Internet Measurement and Data Analysis (14)

Kenjiro Cho

2013-01-09

review of previous class

Class 13 Scalable measurement and analysis (12/26)

- Distributed parallel processing
- Cloud computing technology
- MapReduce
- exercise: MapReduce algorithm

today's topics

Class 14 Privacy Issues

- Internet data analysis and privacy issues
- Summary of the class

privacy

from webster,

- privacy: "the quality or state of being apart from company or observation"
- right to privacy: "freedom from intrusion"

- views on privacy heavily depend on context and culture
  - basic human right
  - a commodity (asset); if infringed, it should be compensated via tort laws (for negligence, liability, etc)

informational privacy

- on collecting, storing and sharing one's personal information

- secrecy of correspondence
  - a fundamental legal principle in the constitutions of many countries
    - against censorship by govenment or any other third party
- naturally extended to communication
  - e.g., telephony, the Internet
- exceptions
  - informed concent (e.g., virus checker/spam filtering services)
  - emergency situations (to protect other services)
  - lawful business practices (e.g., looking at destination in a packet header)
- lawful interception (wire-tapping)
  - suspision of crime
- communication data retention
  - many countries have laws to require service providers to store communication records for certain periods

traceable records everywhere

- cash cards, credit cards, transportatin cards, members cards
- medical records
- device IDs: phone sim, MAC addresses, IP addresses, RFID
- web cookies, geo-location info
- surveillance cameras, fingerprints and machine recognition

your privacy data

- name, date-of-birth, sex, marital status
- address, phone number
- names of family members and pets
- financial data: income, savings, stocks
- educational records, medical records, religion
- purchase data, trip data
- photographs
- online behaviors
- personal preferences
- communication records (when, where, who, how)
- friends

who has your privacy data

authorities

- government bodies
- hospitals
- banks
- universities

commercial services

- stores and other services
- social network services

marketing values

- demographic data, geographic data, other statistics
- condition: individials are anonymous
- existence of black markets (for data thefts)

technological evolutions

technologies increase risks of privacy breaches

- computing power
- database
- consolidation (smartphones, RFIDs)

security and safety vs. privacy

- digital copyright (copyright holders/licensing agencies/users)
- war against terrorism, criminals

- convenience vs. privacy
  - consolidation
  - social networks (your privacy relies on your friends)

privacy in the future

- the post privacy era?
  - we may not have privacy as in the current form in the future
  - the concept of privacy is fairly new
    - since around 1890 after the advent of mass media
- complex issues (cultural, legal, economical aspects)
- as a user, you need to protect your privacy by yourself
  - don't need to be too pessimistic
  - awareness and understanding

previous exercise: WordCount in Ruby

MapReduce-style programming in Ruby

```
% 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
```

on the final report

- select A or B
  - A. PageRank computation of Wikipedia
  - B. free topic
- up to 8 pages in the PDF format
- submission via SFC-SFS by 2013-01-25 (Fri) 23:59

final report topics

A. PageRank computation of Wikipedia

- data: link data within Wikipedia English version (5.7M pages)
- A-1 investigate the distribution of pages
  - A-1-1 plot CDF and CCDF of the outdegree of pages
  - A-1-2 discussion on the outdegree distribution of Wikipedia pages
- A-2 PageRank computation
  - A-2-1 compute PageRank, and show the top 30 of the results
  - A-2-2 other analysis (optional)
  - A-2-3 discussion on the results

B. free topic

- select a topic by yourself
- the topic is not necessarily on networking
- but the report should include some form of data analysis and discussion about data and results

note: you may work with a classmate on programming. but, if you work with someone, make it clear in the report. still, you must write discussions by yourself.

A. PageRank computation of Wikipedia

data: link data of Wikipedia English version (5.7M pages)

- created by Henry Haselgrove (http://haselgrove.id.au/wikipedia.htm)
  - a local copy is avaialble from the class web page
  - a test data set (a subset of 100K pages)
- links-simple-sorted.zip: link data (323MB compressed, 1GB uncompressed)
  - each page has an unique integer ID
  - format: $from : to_1, to_2, ...to_n$
- titles-sorted.zip: title data (28MB compressed, 106MB uncompresed)
  - $n-$th line: the title of page ID n (1 origin)

```
% head -3 links-simple-sorted.txt
1: 1664968
2: 3 747213 1664968 1691047 4095634 5535664
3: 9 77935 79583 84707 564578 594898 681805 681886 835470 ...
%
% sed -n '2713439p' titles-sorted.txt
Keio-Gijuku_University
```

A-1 investigate the distribution of pages

A-1 investigate the distribution of pages

- A-1-1 plot CDF and CCDF of the outdegree of pages
  - include pages with outdegree 0
- A-1-2 discussion on the outdegree distribution of Wikipedia pages
  - optional other analysis
  - hint: you may compare low-degree pages and high-degree pages

A-2 PageRank computation

A-2 PageRank computation

- A-2-1 compute PageRank, and show top 30 of the results
  - format: rank PageRank_value page_ID page_title
  - you may use the script for the exercise
    - use damping factor:0.85 thresh:0.000001
  - takes 5 hours with iMac with 8GB memory (requiring at least 4GB memory)
- A-2-2 other analysis (optional)
  - examples:
  - how to reduce the processing time
  - implement an improved verion of the PageRank algorithm
- A-2-3 discussion on the results

class overview

It becomes possible to access a huge amount of diverse data through the Internet. It allows us to obtain new knowledge and create new services, leading to an innovation called "Big Data" or "Collective Intelligence". In order to understand such data and use it as a tool, one needs to have a good understanding of the technical background in statistics, machine learning, and computer network systems.

In this class, you will learn about the overview of large-scale data analysis on the Internet, and basic skills to obtain new knowledge from massive information for the forthcoming information society.

class overview (cont'd)

Theme, Goals, Methods

In this class, you will learn about data collection and data analysis methods on the Internet, to obtain knowledge and understanding of networking technologies and large-scale data analysis.

Each class will provide specific topics where you will learn the technologies and the theories behind the technologies. In addition to the lectures, each class includes programming exercises to obtain data analysis skills through the exercises.

Prerequisites

The prerequisites for the class are basic programming skills and basic knowledge about statistics.

In the exercises and assignments, you will need to write programs to process large data sets, using the Ruby scripting language and the Gnuplot plotting tool. To understand the theoretical aspects, you will need basic knowledge about algebra and statistics. However, the focus of the class is to understand how mathematics is used for engineering applications.

lessons learned from Internet measurement research

- data collection
  - quality and integrity of data
  - trust/relationship with data owners and users
- data sharing
  - third-party verification: fundamental to science
  - beneficial for society
- privacy in data
  - technical, legal, moral issues
  - social benefits vs. privacy risks

Should we take advantage of big data?

- pros:
  - it helps to convince people for the need of data
  - it attracts researchers, students, and money
  - many useful tools have been developed
- cons:
  - it's just a hype, technically nothing new
  - dubious about those who jump on the bandwagon
  - How can we make use of the big data trend?

big data by cloud services

- “big data” becomes a trendy word, especially for marketing
- most technologies are not new
  - have been used in search ranking, online recommender systems, etc.
- big data processing used to be limited to big organizations that could collect, manage, and analyze data in-house
- now, anyone can easily use big data with cloud services
- package tools are available for collecting and analyzing online customer behaviors
- customer information can be easily used for marketing with minimal initial investment

age of data

- big data is not just for marketing
- technological innovations known as the data revolution are occurring in every field
- previously difficult applications become possible
  - access to huge amount of data, analysis of data constantly being updated, and applications to non-linear models
- big data analysis becomes an indispensable research method in all areas of science and technology

example: impact to science

- e-science: paradigm shift?
  - theory
  - experiment
  - simulation (enabled by computer)
  - data-driven discovery (enabled by big data)

- the supper computing community starts to realize that they should invest for data-sharing rather than computation-sharing

big data technologies

- data collection
  - increasing data sources (e.g., sensors, social media)
- data storage
  - distributed storage, NoSQL database
- data processing
  - distributed/cloud computing (e.g., MapReduce)
- data understanding
  - data mining, machine learning, statistical analysis

big data computing

- use of cloud services
  - utility computing: pay-per-use model
- example: Amazon EC2
  - part of Amazon Web Seervices (AWS)
  - VM (linux or windows, from $0.02/hour)
  - data transfer (from $0.12/GB for download)

computation models

- MapReduce: parallel batch processing
  - e.g., Hadoop, Facebook Puma/Ptail
- Bulk Synchronous Parallel (BSP)
  - e.g., Google Pregel, Apache Giraffe/Hama
- Complex Event Processing (CEP): realtime parallel processing for stream data
  - e.g., Twitter Storm, Yahoo! S4

storage/database systems

- advances in disk technologies
  - capacity
  - access time
  - use of nonvolatile solid-state memory (SSD, etc)
- new perspectives (for data analysis)
  - most measurement data is write-once
    - consistency/locking can be relaxed
    - large block size
  - most data types are simple: e.g., key-value

file systems

- distributed file systems
  - fault-tolerance
  - scalability
  - huge file support
  - e.g., GFS, HFS, GlusterFS, MogileFS, NFSv4.1

NoSQL database

- key-value store: (simple key-value type)
  - e.g., Dynamo, Redis, Voldemorte, Membase
- column-oriented DB: (optimized for column access)
  - e.g., BigTable, Hbase, Cassandra
- document oriented DB: (schemaless)
  - e.g., MongoDB, CouchDB, SimpleDB
- graph-oriented DB: (index-free adjacency)
  - e.g., InfoGrid, Neo4j

UNIX shell and pipe

- old technology, but still most useful
  - becomes much more efficient on multicore machines
  - most of data analysis can be done on a single machine
    - with appropriate pre-processing

- my favorite and most often used method!

data analysis is merely a tool

- recent big data trends focus too much on tools and methods but data analysis is merely a tool
- data analysis is an iterative process
    - forming a hypothesis, verifying it with data
    - if the results are unexpected, you find new questions
    - repeating the process will uncover interesting facts
- analysis without purpose ends up with useless numbers
- If you identify what to get from data, you will see a path forward

privacy issues

- increasing risk of privacy breaches by big data and data mining
  - anonymization and de-identification of personal info
  - personal info: health, location, electricity use, online activity
- there are technical, legal, moral aspects
- hard to assess risks in the future

issues in the data age

- shortage of data specialists
  - needs of data scientists with field-specific knowledge and data analysis skills
  - challenge conventional thinking and interpretations, establish issues clearly, use statistics and data as tools to resolve problems
- data as assets
  - if based on the same data, success depends on analytical ability
  - but if data quality varies, big advantage with better quality data

issues in the data age (cont 'd)

- data sharing
  - social benefits of data sharing and third-party verification
  - a balance of social benefit and risk of privacy breaches
- social consensus making
  - how far private/public organizations can track individuals
  - how information (e.g., personal medical records) will be shared for social benefit

recipient literacy

- info recipients also need to understand and question data
  - as there are more and more questionable data and dubious theories based on data
- we tend to want to see things as either black or white
  - but most things are grey; determining black/white is to draw a line just for convenience
  - seeking a black or white answer is to avoid own judgment and responsibility
- we should accept grey as grey, and make our own judgement

fundamental change to creative thinking process?

- data-driven decision making has been always important
- but, ICT pushes it to a completely different level (in quality, quantity, expressions)
- now, we can literally interact with data (data-human interaction)

big data summary

- big data is a hype, still some part is useful for Internet measurement and data analysis
  - to attract people and money
  - useful tools
- increasing importance of data
  - it will change how we think
  - our experiences from Internet measurement would be useful for others

summary

Class 14 Privacy Issues

- Internet data analysis and privacy issues
- Summary of the class

summary of the class

- Internet measurement and data analysis
- large-scale data analysis on the Internet
- learned technical background using programming