Human Interaction with Data

Kai Shen

Human Interaction

- People are consumers of data
  - Data processing results are to improve human knowledge and decision-making
  - Visualization
- People are sometimes the data sources
  - Wearable sensors and smart phones provide rich data on people
- Human efforts help big data tasks
  - Assist data collection, processing,analysis
  - Crowdsourcing

Big Data Visualization

- Illustrate the data in a graphic way that is easier for people to make sense, find insights
- Large data volume
  - One pixel per record
    "Visual exploration of large data sets", Keim, 2001, Figure 1
  - Treemap to show structural information
    "Interactive Information Visualization of a Million Items", Fekete and Plaisant, Figure 1
- Many data dimensions
  - Dimension reduction (e.g., principal component analysis)
  - Visualization for patterns, clustering

Interactive Visualization

- Typical human exploratory process
  - Overview first
  - Zoom and filter
  - Find details on demand

⇨ Dynamic, interactive visualization that allows zooming, marker highlighting/filtering, …

Tag Clouds

- Visualization for texts
  - Tags are usually single words
  - Importance of each tag is shown with font size
  - http://en.wikipedia.org/wiki/File:Web_2.0_Map.svg
- Perception of tag clouds [Lohmann et al. 2009]
  - Tag size:
  - Position (center vs. border; upper left vs. lower right)
  - Users tend to scan tag clouds

Human-Assisted Data Collection

- Smart phones carry many sensors
  - GPS, accelerometers, microphone, cameras, …
- Collect data on people themselves
  - Data on daily movement to help health and wellness
  - Data on user whereabouts to tell user interests and activities
  - User facial expression analysis (mood meter) to label web content and enable adaptively interactive applications

  ⇨ Large business value for ads, but must address privacy concerns

Human-Assisted Data Collection

- Crowdsourcing data acquisition:
  - Leverage human activities to collect data
- Smart phone power usage characterization
  - Modern smart phone power consumption varies over a wide range, depending on the human uses and applications
  - To understand typical smart phone power consumption and its correlation with certain uses and applications

  ⇨ Very helpful if many users are willing to send data on their phone uses, applications, and power to us

Human-Assisted Data Collection

- Traffic monitoring
- Existing mechanism
  - Tracking devices on trucks that show truck locations and conditions
- Mobile Millennium project at UC Berkeley
  - Software running on mobile phone that detects location and speed
  - Many users collect such data on their phones, centralize the data for analysis and filtering, to derive traffic condition information

Human-Assisted Data Collection

- Novel use of existing data on user activities
- Why Google can build a better spellchecker than Webster dictionary does?
  - Has hit counts on alternative spellings
  - More importantly, has data on how often suggested spellings are adopted by uers

Human-Assisted Data Collection

- http://www.google.org/flutrends/
  - Infer Flu trends and locations through user searches
  - Search keywords suggest flu events
  - IP addresses of searchers (and keywords sometimes) tell the locations
- May do better with Tweeter or Facebook data.

User-Generated Content Sites

- Wikipedia
- WeBWorK
- Users contribute the content, collaboratively edit and curate the content

Human-Assisted Data Processing

- Human do certain things better than computers do
- Labeling images
  - “Games with a purpose”, von Ahn and Dabbish, 2008
- Optical character recognition
  - Useful in digitize books or newspapers that are not yet digitized
  - Re-captcha: two CAPTCHAs, one to verify a human user and the other to ask user solution on optical character recognition that computers have so far failed
    - Win-win for Google and web sites who need CAPTCHA
    - Need to make the two CAPTCHAs hard to distinguish

Amazon Mechanical Turk

- Besides image labeling and optical character recognition, there are many other things humans can do better
  - Find "duplicate" web pages, identify spam sites, rewrite a sentence, fill in a survey, comment on a blog, vote for a tweet or an online poll, …
- Amazon Mechanical Turk: an online marketplace for buying/selling human effort and intelligence
  - Programmable interface for the buyers

Human Uses in Data Computation and Management

- TurKit [Little et al. 2010]
  - Figure 1
  - Figure 2
  - Human computation is expensive, must optimize to conserve it in case of crashes
- CrowdDB [Franklin et al. 2011]

  SELECT profile FROM department

  WHERE name CROWDEQUAL "CS"

Implications on Computer Systems

- Parallel computing, GPU
- Mobile computing, Power
- Programming languages
- Databases
- … …

Disclaimer

- Preparation of this class was helped by materials in the book "Frontiers in Massive Data Analysis" by the Committee on the Analysis of Massive Data, the Committee on Applied and Theoretical Statistics, the Board on Mathematical Sciences and Their Applications, the Division on Engineering and Physical Sciences, and the National Research Council.