Big Data Collection

Kai Shen

Challenges of Big Data Collection

- Challenging to acquire a lot of data quickly
  - Need of large processing, networking, and storage throughput
  - Parallelism can help
- Challenging to acquire useful information from a sea of irrelevant data
  - What data is more important than others
  - Identify redundancy efficiently
  - Collect topic-specific data
- Challenging to collect from distributed, remote data sources

Web Crawling

- Collect published web content – crawling
  - First retrieve some root/seed pages;
  - Parse their content and follow hyperlinks to retrieve more pages;
  - Repeat the last step.
- How is it useful?
  - Web search engines
  - Dive deep (go beyond what is provided by search engines) on a specific topic
  - Businesses/advertisers to find potential customers
  - … …

Goals

- Collect good (high-quality) pages
- Collect web pages on a certain topic
- Crawl efficiently
- Crawl without annoying others

Redundancy Removal

- Avoid parsing and following the same page more than once
  - Record all URLs that have been parsed and followed
  - Same page may have different URLs
- URL normalization
  - Host portion is case-insensitive
  - Decode percent-encoded octets of unreserved characters ('%30' is '0')
  - '/./' or '//'
  - … …
- URLs that look totally irrelevant may also be the same page
  - Compute and match page content checksum

Link Selection

- There are many choices at each step of crawling
- Depth-first search vs. breadth-first search?
- Hyperlink with high likelihood pointing to a high-quality page?
  - High in-links
  - Pointed to from high-quality pages
  - Spam identification and avoidance
  - … …

Focused Crawling

- Maximize the page downloads on a certain target, minimize the resources spent on irrelevant downloads
- Crawl all pages from a domain (e.g., rochester.edu)
- Topic-specific crawling:
  - A page's topic is inferred from the URL text, anchor text, and surrounding text of the hyperlink
  - Look for specific keywords
  - Reinforcement learning [McCallum, Rennie, and others 1999]
    - Train a Q function: mapping from "bag of words" to a value (future rewards – chance of hitting on-topic pages if following the link)

Scalable Web Crawling

- What are the resources consumed?
  - CPU processing for network operations and the parsing of page content
  - writing to disk storage
  - network bandwidth to remote web sites
- Parallel web crawling
  - Use multiple CPUs
  - Use multiple storage devices
- Challenges:
  - Synchronization on already visited URLs
  - Downloading network bandwidth remains the bottleneck

Politeness

- Crawling is not always welcome (in fact, often unwelcome) by web sites
  - Some of my content is good to show up on a search engine, others (processing scripts etc.) has no use (to me) to be crawled
  - Crawlers/bots consume my web server resources that may affect the experience of human users
- Robot Exclusion Standard
  - http://www.cs.rochester.edu/robots.txt
- Limit the crawler bandwidth use
  - Page download rate; downloading bandwidth
  - Limit per site

Assignment #1

- Focused web crawling
  - Within a domain
  - Topic-specific
- Not to visit the same URL twice, URL normalization
- Follow the robot exclusion standard and crawl at a determined slow rate (e.g., sleep a second between consecutive page downloads)
- No parallelism

Assignment #1

- Parsing a web page for hyperlinks and anchor texts?
  - HTML pages are notoriously error-prone ⇨ browsers go to great length to tolerate imperfect HTML pages; you may find it not easy to do
- Maintaining the list of visited URLs ⇨ what data structure?
- Programming languages:
  - Java, C, Perl, Python, …
- Demo and pre-assignment questionnaire

Other Big Data Collection

- USArray seismic data collection
  - http://www.usarray.org/
- Seismic data collection in African volcanos [Prof. Ebinger at the Earth and Environmental Sciences]

Other Big Data Collection

- Challenges of data collection at remote locations
  - No power infrastructure ⇨ solar panels + batteries
  - Reliability (batteries died; GPS malfunctioning)

Other Big Data Collection

- High-speed cameras on roads and highways
- Challenges
  - A lot of data produced by high-speed cameras
  - Wireless deployment is economically efficient, but wireless networks aren't very good (low bandwidth, intermittent)
  - Processing on site to relieve the burden of data transfer

Public Eavesdropping

- Listen to open WiFi signals
  - Lots of communications are unencrypted
  - WEP encryption is weak
- Insert a Gnutella relay node to collect information on Gnutella traffic
  - Know what are being searched
  - Know the group of searches made by one person

Big Data in Your House

- If you are Google, Facebook, Twitter, Amazon, AT&T, VISA, Strong Hospital, …
  - There are limited privacy laws
  - Others are fair game (individual identities are usually scrambled in the collection phase)