B345 Internet Science and Technology

Week 11 lecture 2

Today's Lecture Learning Objective

- Understand the techniques for
  - Gathering web measurement data
  - Processing measurement data
  - Deriving workload models from measurement data

Measurement Techniques

- Where to get data about what is happening on the web.
    - Server logs
    - Client logs
    - Proxy logs
    - Packet monitoring
    - Active measurements
    - User-centric measurements

<< See Diagram A & B >>

Server Logs

- Logs done by web servers - information about web site access.
- Analysis tools.
- Issues:
  - Can't log much details.
  - Cached responses.
  - Can't identifying users - proxies, shared clients, dynamic IP, etc.
  - Web sites vary.

Log File Formats

- Common Log Format (CLF)
  - Remote host
  - Remote identity
  - Authenticated user
  - Time
  - Request
  - Response code
  - Content length
- Extended Common Log Format (ECLF)
- Individual web server's formats

Client Logs

- Logs on browsers - information about users.
- Possible to log everything.
- Issues:
    - No standard log formats.
    - Controlling which browsers.

Proxy Logs

- Client or server-side proxies.
- Standard log formats for server proxies.
- Advantages and disadvantages compared to server and client logs.

Packet Monitoring

- Monitoring TCP/IP traffic
  - Useful connection information
  - Eases burden on web components
- Issues:
  - Easy for some lower level technologies, hard for others.
  - A lot of traffic.
  - Routing not deterministic.

Active Measurements

- Generate requests and test - controlled data.

- Issues:
  - Where to locate user agent.
  - What request to generate.
  - What measurements to collect.

User-centric Measurements

- A panel of users to report on web behaviours.

- Issues:
  - Representativeness of panel.
  - Misreporting.

Processing Measurement Data

- What to do with measurement data once we have got it.
    - Parsing
    - Filtering
    - Transforming
    - Analysing

Parsing Data

- Identifying the fields in a particular log file.
- Reading in the relevant values.

Filtering Data

- Removing irrelevant fields or entries.
- Criteria for what is relevant and what is not.

Transforming Data

- Changing data formats to something easier to analyse:
    - Message lines
    - Time formats
    - IP addresses

Analysing Data

- Some information direct from data.
    - Eg. traffic volume, number of requests, etc.
- Others have to be inferred:
    - HTTP headers
    - Ambiguous identity
    - User actions
    - Resource modifications

Characterizing Web Workloads

- Making up statistical models that describe web characteristics.

- Used for
  - Identifying performance problems
  - Benchmarking web software and hardware
  - Capacity planning

Web Characteristics

- HTTP messages
  - What Request Methods
  - What Response Codes
- Resources
  - Content Types
  - Resource Size
  - Response Size
  - Popularity
  - Modification frequency
  - Temporal locality

Web Characteristics

- User behaviour
  - Session interarrival times
  - Number of clicks per session
  - Request interarrival times

Applying Workload Models

- Benefit of an accurate workload model.
- Using the workloads models to test.