

Data Mining

Web Mining I – Web Usage Mining (Part B)

Dr. Jason T.L. Wang, Professor
Department of Computer Science
New Jersey Institute of Technology

Web Usage Mining

Four processing stages of Web usage mining:

- usage data collection,
- usage data preprocessing,
- usage pattern discovery,
- usage pattern analysis (e.g. construct a Web log data cube and apply OLAP operations).

Web Usage Data Collection

Usage information can be collected from the following sources:

Web server – stores access logs to the server

Web proxy server – stores access logs from anonymous users
sharing the same proxy server

Client machine – stores browsing logs on the client side

Web Server Log Data

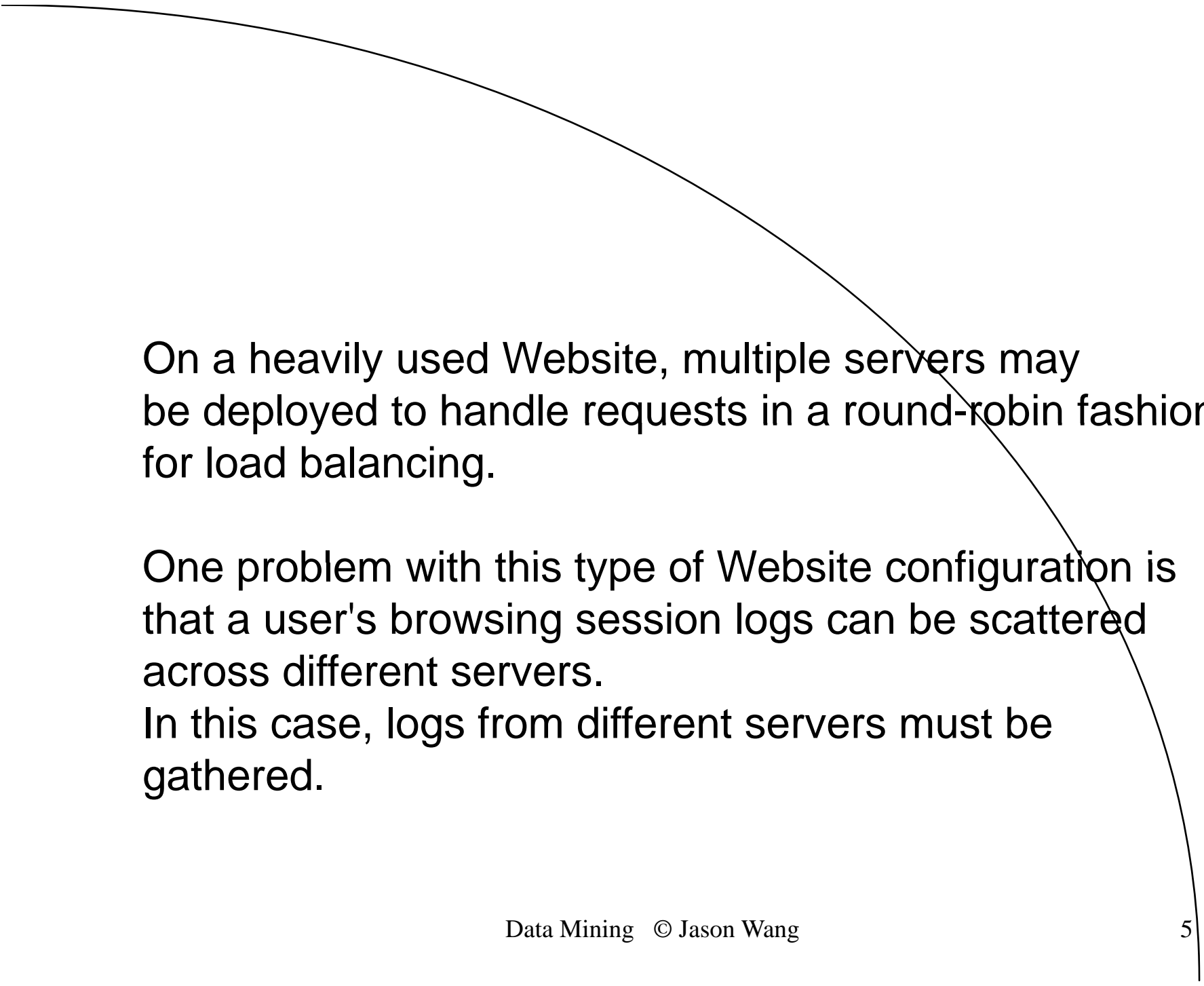
Web servers store access request information in Web server access logs.

Access logs are like fingerprints characterizing Web servers. For each browsing session to a Web server, entries are recorded in the following log files:

Access logs – store client access information (date, client IP, request URL, bytes transferred, etc.)

Error logs – store failure client access information

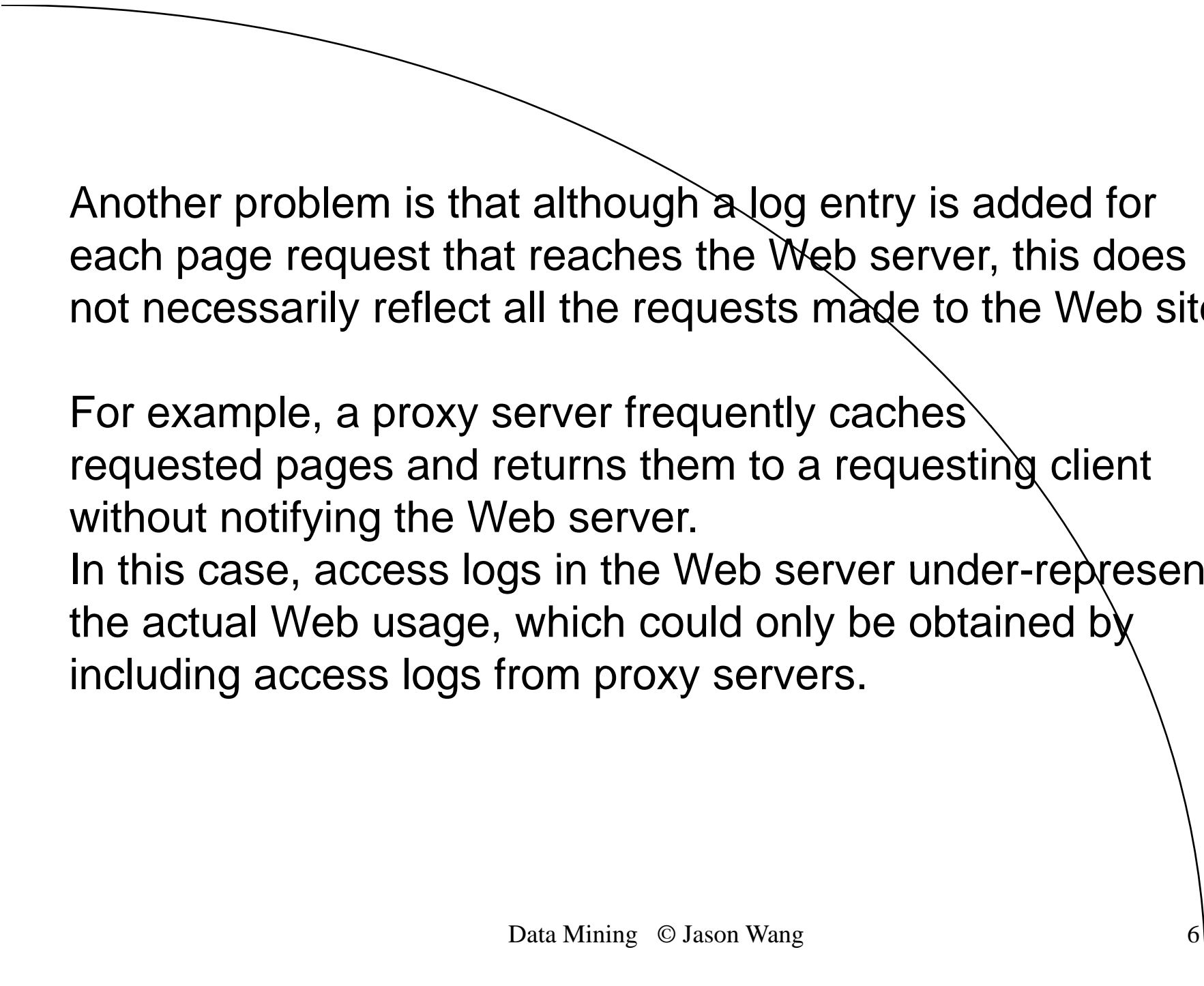
Cookie logs – store session information between client and server



On a heavily used Website, multiple servers may be deployed to handle requests in a round-robin fashion for load balancing.

One problem with this type of Website configuration is that a user's browsing session logs can be scattered across different servers.

In this case, logs from different servers must be gathered.



Another problem is that although a log entry is added for each page request that reaches the Web server, this does not necessarily reflect all the requests made to the Web site.

For example, a proxy server frequently caches requested pages and returns them to a requesting client without notifying the Web server.

In this case, access logs in the Web server under-represent the actual Web usage, which could only be obtained by including access logs from proxy servers.

User Activities

Page view – visual rendering of a Web page in a specific client environment at a specific point in time

Click stream – a set of user-initiated requests which can be either explicit, implicit, embedded, or user clicks

Server session – a collection of user clicks to a single Web server during a user session (also called a visit)

User session – a delimited set of user clicks across one or more Web servers

Episode – a subset of related user clicks that occur within a user session

Statistical Analysis

Many log analysis tools use this technique to analyze site traffic including frequently accessed pages, average file size, daily traffic, the number of site visitors, access error reporting, etc.

The discovery of facts about a Website is potentially useful for monitoring Web usage, security checking, performance tuning, and site improvement.

Web Usage Mining

Association Rule Discovery:

For example, marketing people can use the discovered association rules for marketing applications while Web masters can use them for site restructuring.

Clustering:

For example, in clustering user sessions, customers having a similar browsing pattern are grouped together.

Web Usage Mining

Classification:

Given a user's browsing pattern, a classification technique can be used to classify this user into different categories of interests.

Sequential Pattern Discovery:

For example, certain users might access certain pages with periodicity. Periodic patterns can be discovered through this type of analysis, which is useful in discovering trends in databases.

Techniques for Web Usage Mining (A Summary)

- Construct multidimensional view on the Weblog database.
 - Perform multidimensional OLAP analysis to find the top N users, top N accessed Web pages, most frequently accessed time periods, etc.
- Perform data mining on Weblog records.
 - Find association patterns, sequential patterns, and trends of Web accessing.
 - May need additional information, e.g., user browsing sequences of the Web pages in the Web server buffer.
- Conduct studies to
 - Analyze system performance, improve system design by Web caching, Web page prefetching, and Web page swapping.



End of Web Usage Mining Module (Part B)