

Data Mining

Web Mining II – Web Structure Mining (Part A)

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

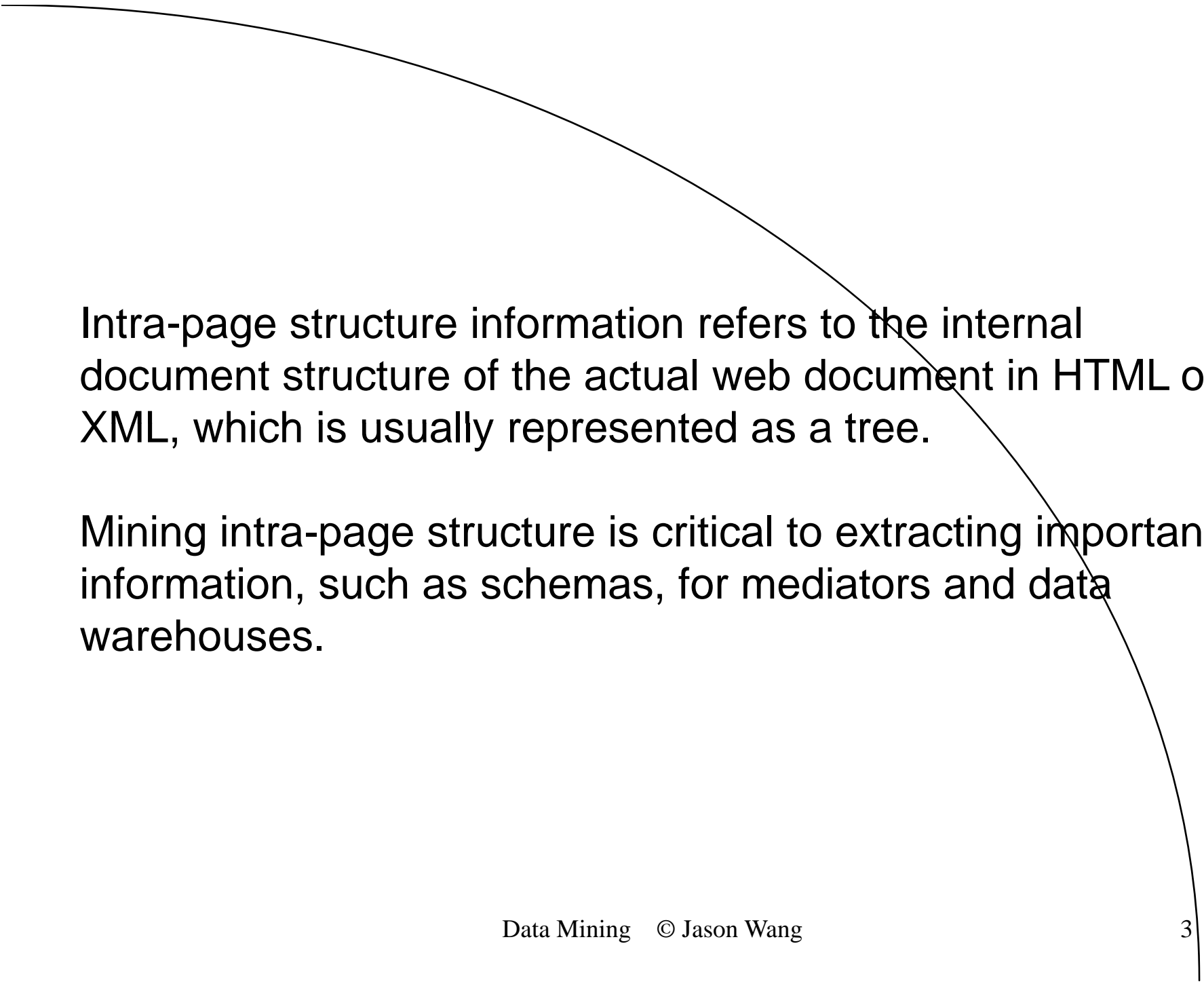
Web Structure Mining

Web structure mining is the process of analyzing the structured information used to describe Web content. Structured information on the Web can be broadly classified as intra-page or inter-page.

Inter-page structure information can be analyzed by traversing hyperlinks, and is often called Web linking structure.

It is a rich source of information about the nature of the Web. In this type of mining, the linking structure can be represented as a graph in which Web documents are the nodes and hyperlinks are the directed edges of the graph.

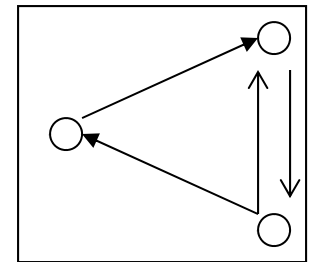
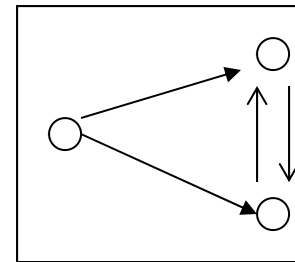
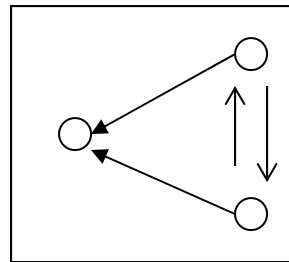
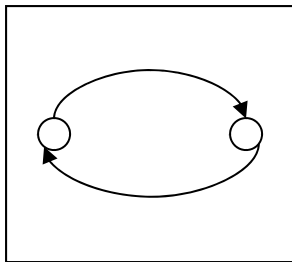
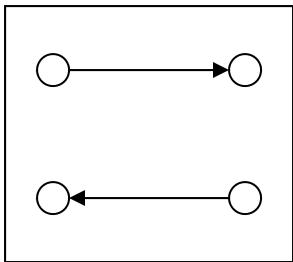
Useful information can be discovered by processing the relationships between nodes and edges.



Intra-page structure information refers to the internal document structure of the actual web document in HTML or XML, which is usually represented as a tree.

Mining intra-page structure is critical to extracting important information, such as schemas, for mediators and data warehouses.

Basic Hyperlink Relationships

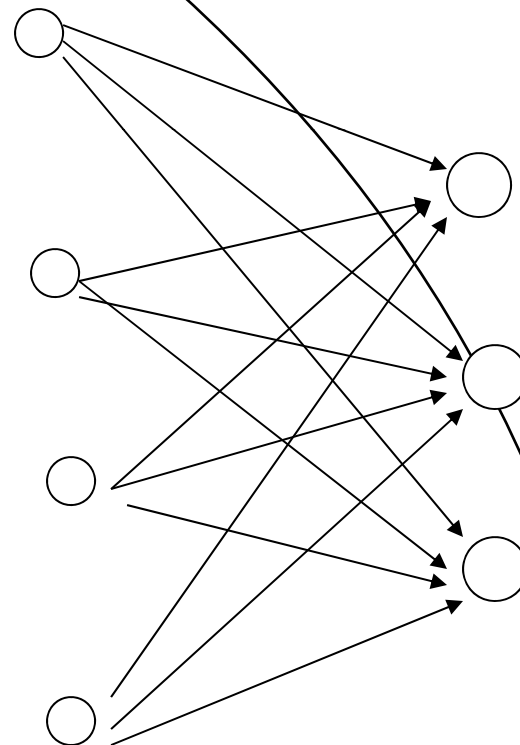
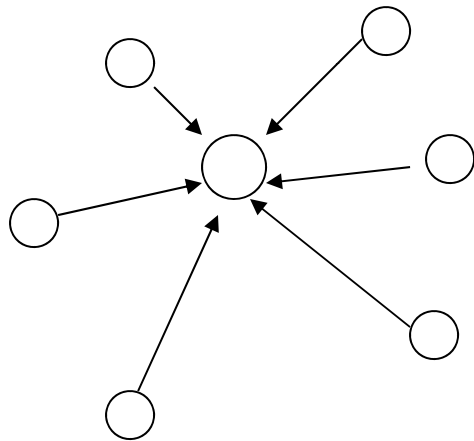


Web Structure Mining

Hubs and authorities exhibit strong mutually reinforcing relationships because a hub becomes a better hub when it links to many good authorities.

Likewise, an authority becomes a better authority when it is linked by many good hubs.

This type of analysis is called *connectivity analysis*.





End of Web Structure Mining Module (Part A)