

# **Data Mining**

## **Query Based Web Search Systems**

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

# Query-Based Web Search Systems

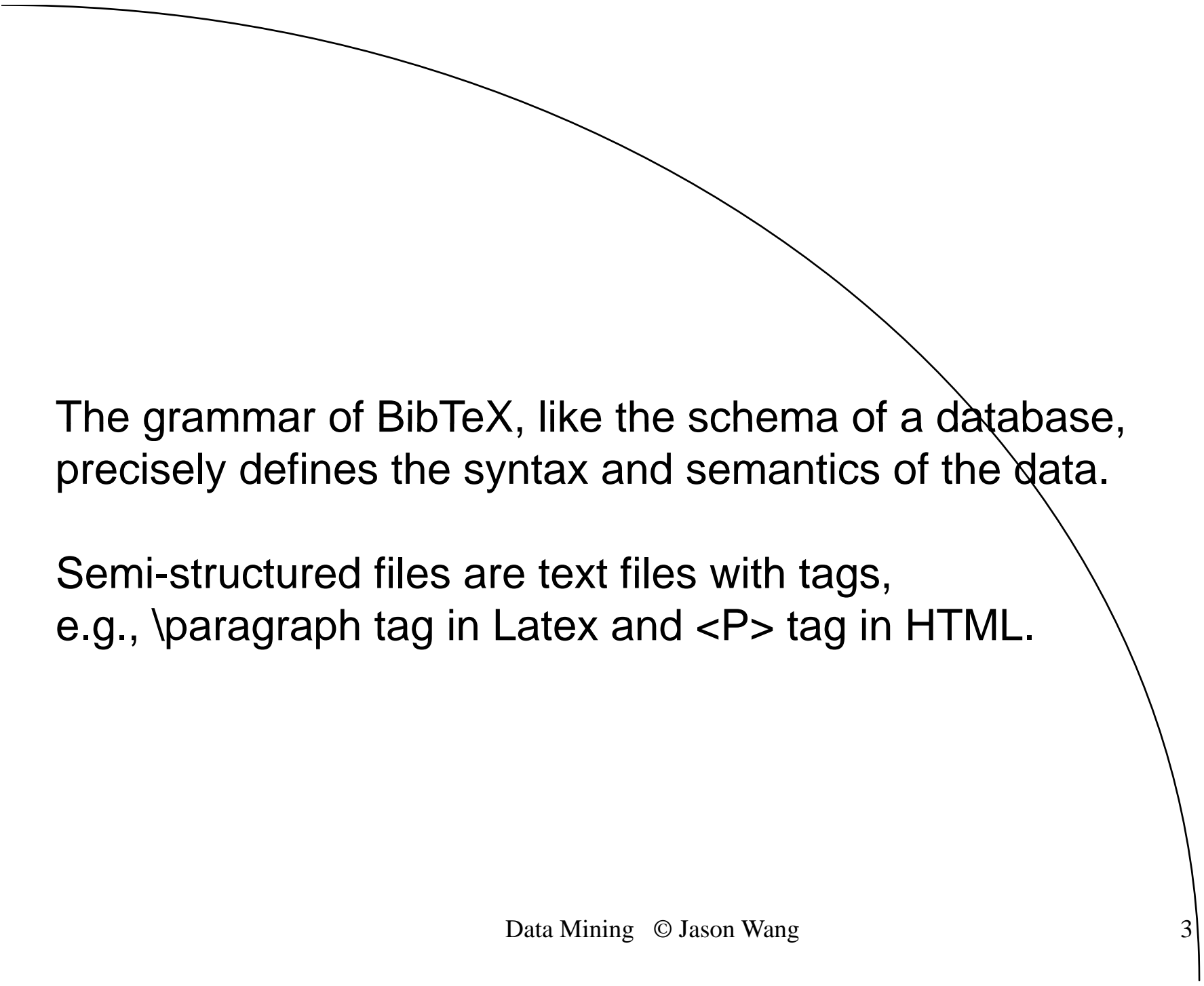
Keyword based search engines lack SQL-like queries.

Systems allowing SQL-like queries include W3QS/W3QL, WebSQL, etc.

They interact with the Web directly.

They need to deal with three types of Web data:

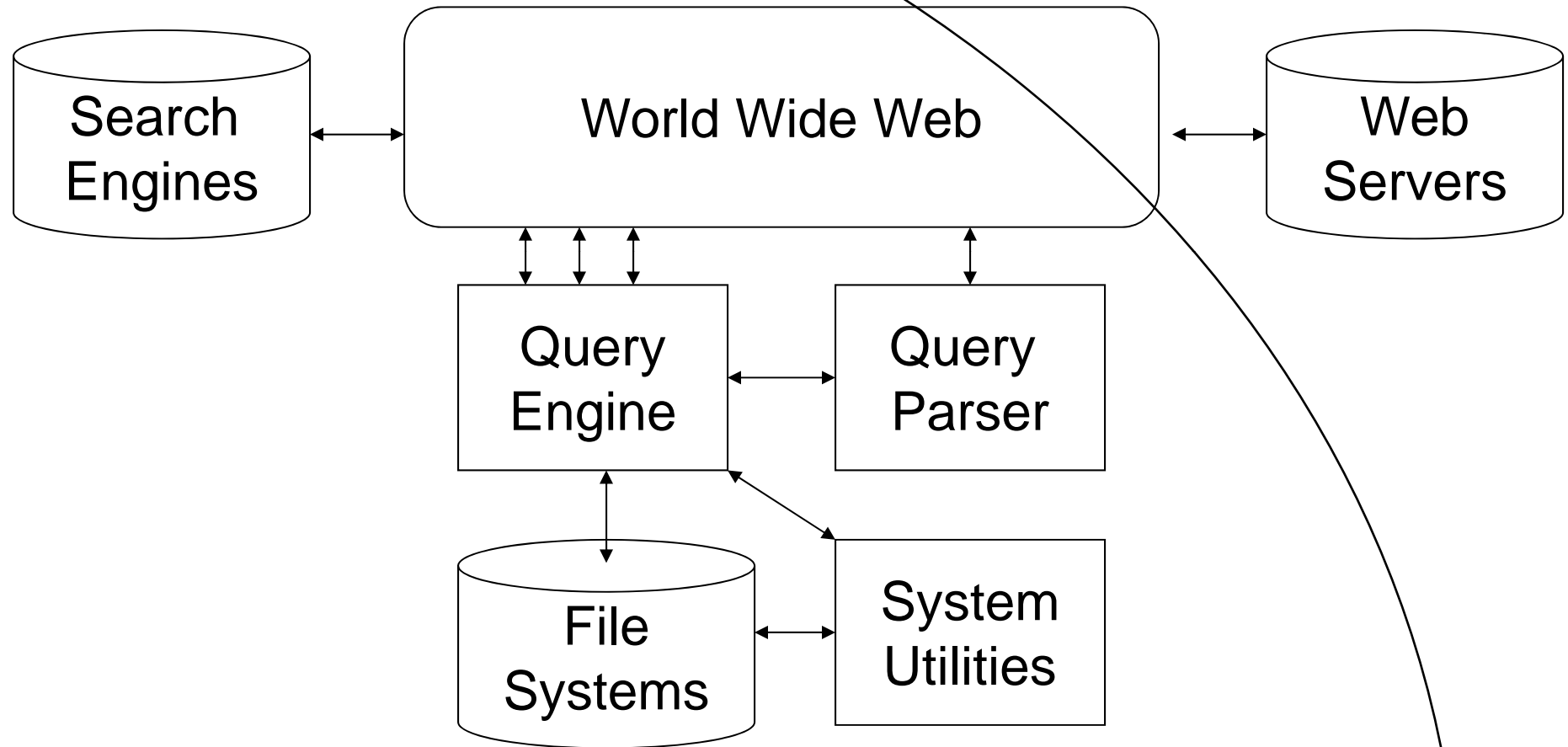
- Structured data, such as flat databases and BibTeX files;
- Semi-structured data, such as HTML, XML, Latex files;
- Unstructured data, such as sound, image, pure text and executable files.



The grammar of BibTeX, like the schema of a database, precisely defines the syntax and semantics of the data.

Semi-structured files are text files with tags, e.g., `\paragraph` tag in Latex and `<P>` tag in HTML.

# Architecture of a Web Querying System



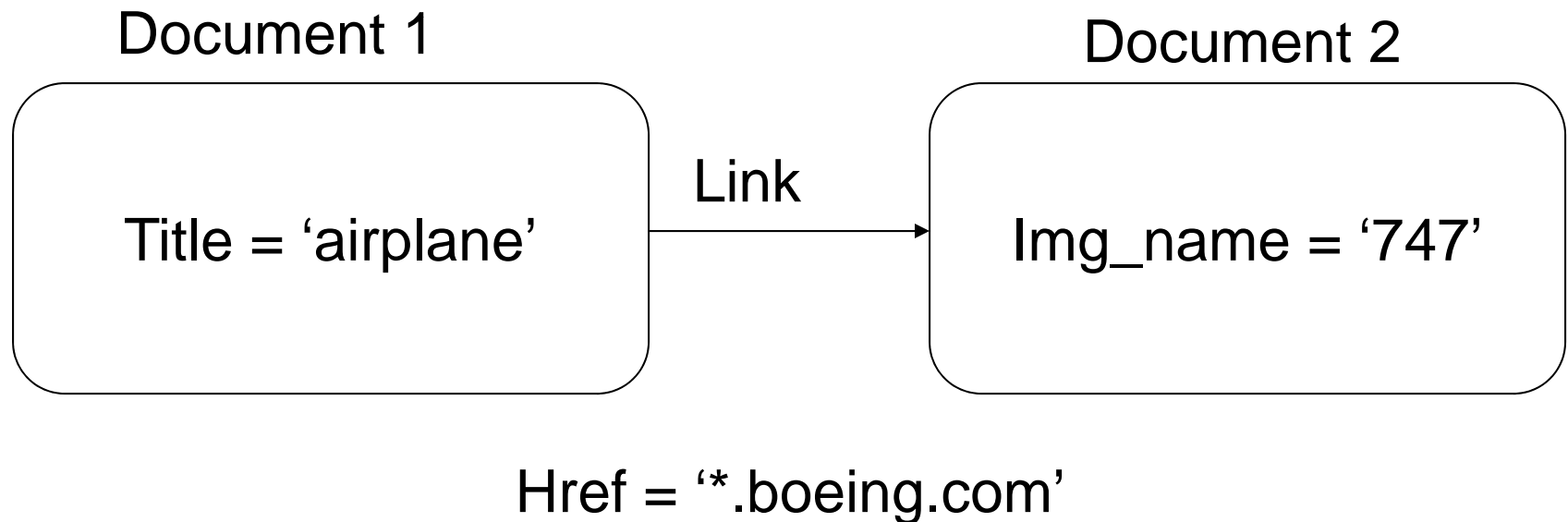
# W3QS/W3QL

W3QL views the Web as a directed graph.

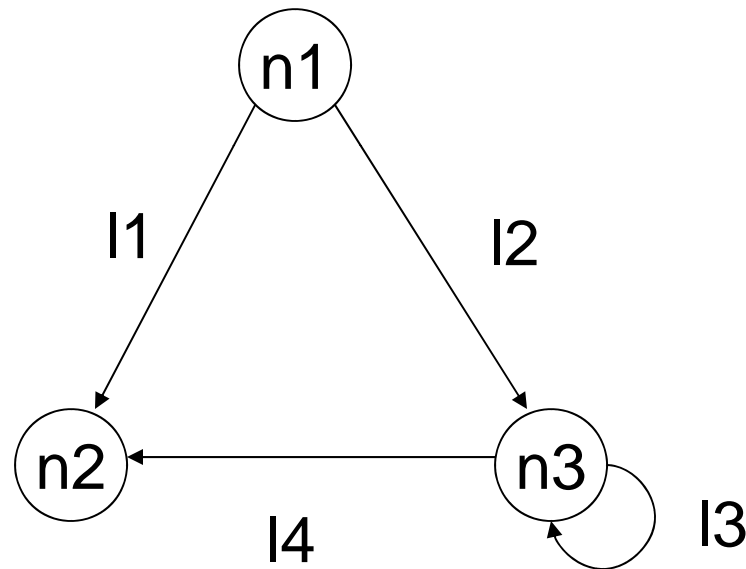
- Each URL or document is a node.
- A directed edge from node x to node y means node x is an HTML or XML document and it contains at least one anchor (hyperlink) to node y.

# Structure-based Query

The following structure-based path pattern query finds documents with "airplane" in their title and with a hyperlink to Boeing's Website with an image of 747.



# A Pattern Graph (Path Expression)



n1 l1 n2

n1 l2 (n3 l3)

(n3 l3) l4 n2

or n1 l2 (n3 l3) l4 n2

# W3QL Example Queries

Example. This query searches for HTML documents that contain hyperlinks to images in GIF format located in www.xyz.com. Here n1 and n2 are nodes and l1 is the link used to connect n1 with n2. The select statement is a command that copies the content of n1 to a file called result.

```
select  cp n1/* result;
```

```
from    n1, l1, n2;
```

```
where   SQLCOND (n1.format = HTML) AND
```

```
(l1.href = "www.xyz.com") AND
```

```
(n2.name = "*.gif");
```



# WebSQL

WebSQL is also based on a graph model of a document network. It views the Web as a “virtual graph” whose nodes are documents and whose directed edges are hyperlinks. To find a document in this graph, one navigates starting from known nodes or with the assistance of index servers.

Given a URL  $u$ , an agent can be used to fetch all nodes reachable from  $u$  by examining anchors in the document contents. Conversely, nodes that have links to  $u$  can be found by querying index servers for nodes that contain links to node  $u$ , using link:url search format in the AltaVista search engine.

Example 1. This query is used to search for HTML documents about “mining”.

It returns the URL of the documents in the tuple d.

```
select d.url
```

```
from Document d
```

```
SUCH THAT d MENTIONS "mining"
```

```
where d.type = "text/html";
```

Example 2. This query is used to search for documents about “mining” that contain a hyperlink to [www.kdd.org](http://www.kdd.org). It returns the URL and title of the documents in the tuple d.

```
select d.url, d.title
from   Document d
      SUCH THAT d MENTIONS "mining", Anchor y
      SUCH THAT base = d
where  y.href = "www.kdd.org";
```

Example 3. This query is used to search for documents with “Web mining” in the title, which are linked from a hyperlink path originating at www.kdd.org, of length two or less, and located on the local server. It returns the URL and title of the documents in the tuple d.

```
select d.url, d.title
from   Document d
  SUCH THAT "http://www.kdd.org" = | --> | --> --> d
where   d.title = "Web mining";
```



# **End of Query Based Web Search Systems Module**