Data Mining

Web Crawling (Part A)

Dr. Jason T.L. Wang, Professor

Department of Computer Science

New Jersey Institute of Technology

Web Crawling Agents

Agents:

• Autonomous agents:- self-maintained programs that can move between hosts according to conditions in the environment (they can travel only between special hosts for security reasons and are not widely used on the Internet).

• Intelligent agents:- programs which assist users in finding requested data items, filling forms, and using software (host-based software that has little to do with networking).

• User-agents:- programs which execute requests for services through a network on behalf of a user (they need constant interaction and have limited intelligence).

For example, Netscape Navigator is a Web user-agent and Eudora is an email user-agent.

• Web crawling agents:- They are a combination of agents and Web crawlers, responsible for intelligently gathering important data from the Web (exhibiting the characteristics of both intelligent agents and user-agents).

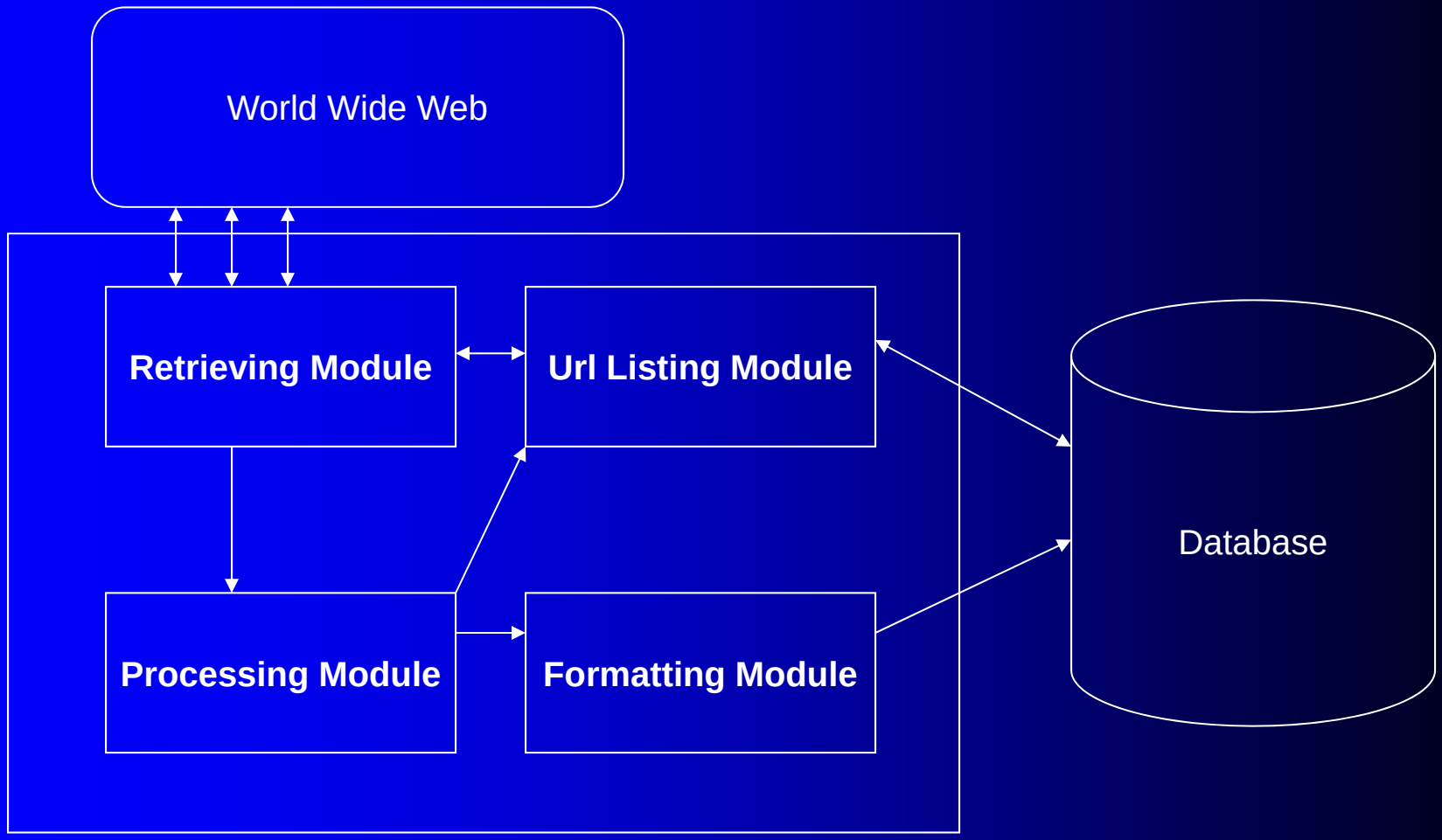
Web crawlers (or crawlers, robots, spiders, walkers, wanderers, worms)

- They are programs, which automatically traverse the Web via hyperlinks embedded in hypertext, news group listings, directory structures or database schemas.
- Crawlers are responsible for gathering resources from the Web, such as HTML documents, images, postscript files, text files and news-postings.
- In contrast, browsers are user-driven, not automatic. Research shows that roughly 190 million Web pages have been indexed by six major search engines. The indexed pages were visited by crawlers.

Tasks performed by crawlers

- Indexing - to build and maintain indexes for search engines.
- HTML validation - to check whether a page conforms to HTML DTD.
- Link validation - to check whether a link is still valid.
- Information monitoring – to monitor changes to HTML pages.
- Information search - to search for required documents.
- Mirroring – to build mirror (duplicate) sites.

Web Crawling Architecture



Web Crawling Algorithms

Most crawlers are unable to visit every possible page for the following reasons:

• Apart from scanning the Web for new pages, a crawler must also periodically revisit pages already seen, to update changes made since the last visit. These changes affect the index and crawling paths.

- Server resources are limited by their storage capacity. The Web has been estimated to contain 320 million indexable pages and has already grown probably beyond one billion (10^9) pages.
- Network bandwidth is limited to the type and number of connections to the Internet. Only a limited number of simultaneous crawlers can be launched by a given Web crawling system.

Smart crawling techniques are needed to capture the most current view of the Web.

Let

V = { p_x, ... ,p_y } be a set of visited pages and

U = { p_m, ... ,p_n } be a set of pages, called front pages, defined as follows:

U consists of unvisited pages only, so U and V are disjointed.

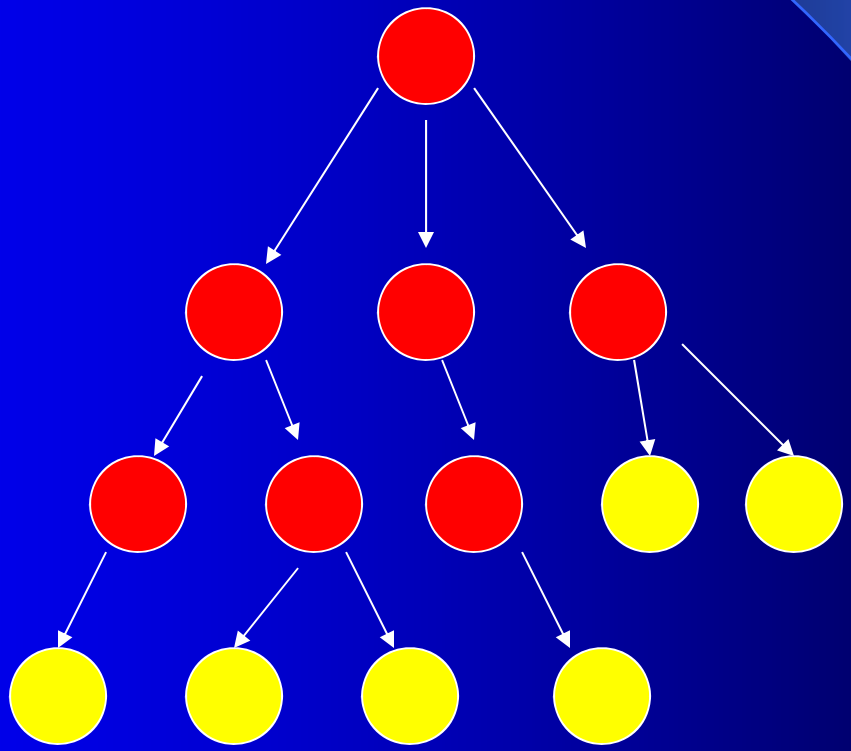
For all p_i in U, there exists a p_j in V, such that

p_i is 1-distance away from p_j (i.e., every page in U is at most one link away from a page in V).

Since hyperlinks are directed paths, the converse may not hold.

At any given instant, the crawler selects “the most important” front page to visit.

Red Nodes: Visited Pages
Yellow Nodes: Front Pages



End of

Web Crawling Module (Part A)