Artificial Intelligence 15-381
Web Spidering & HW1 Preparation

Jaime Carbonell
jgc@cs.cmu.edu
22 January 2002

Today's Agenda

- Finish A*, B*, Macrooperators
- Web Spidering as Search
  - **How to acquire the web in a box**
  - **Graph theory essentials**
  - **Algorithms for web spidering**
  - **Some practical issues**

Search Engines on the Web

Revising the Total IR Scheme

1. Acquire the collection, i.e. all the documents

[*Off-line process*]

2. Create an inverted index (IR lecture, later)

[*Off-line process*]

3. Match queries to documents (IR lecture)

[*On-line process, the actual retrieval*]

4. Present the results to user

[*On-line process: display, summarize*, ...]

Acquiring a Document Collection

Document Collections and Sources

- **Fixed, pre-existing document collection**

  e.g., the classical philosophy works

- **Pre-existing collection with periodic updates**

  e.g., the MEDLINE biomedical collection

- **Streaming data with temporal decay**

  e.g., the Wall-Street financial news feed

- **Distributed proprietary document collections**

  Distributed, linked, publicly-accessible documentse*.g. the Web*