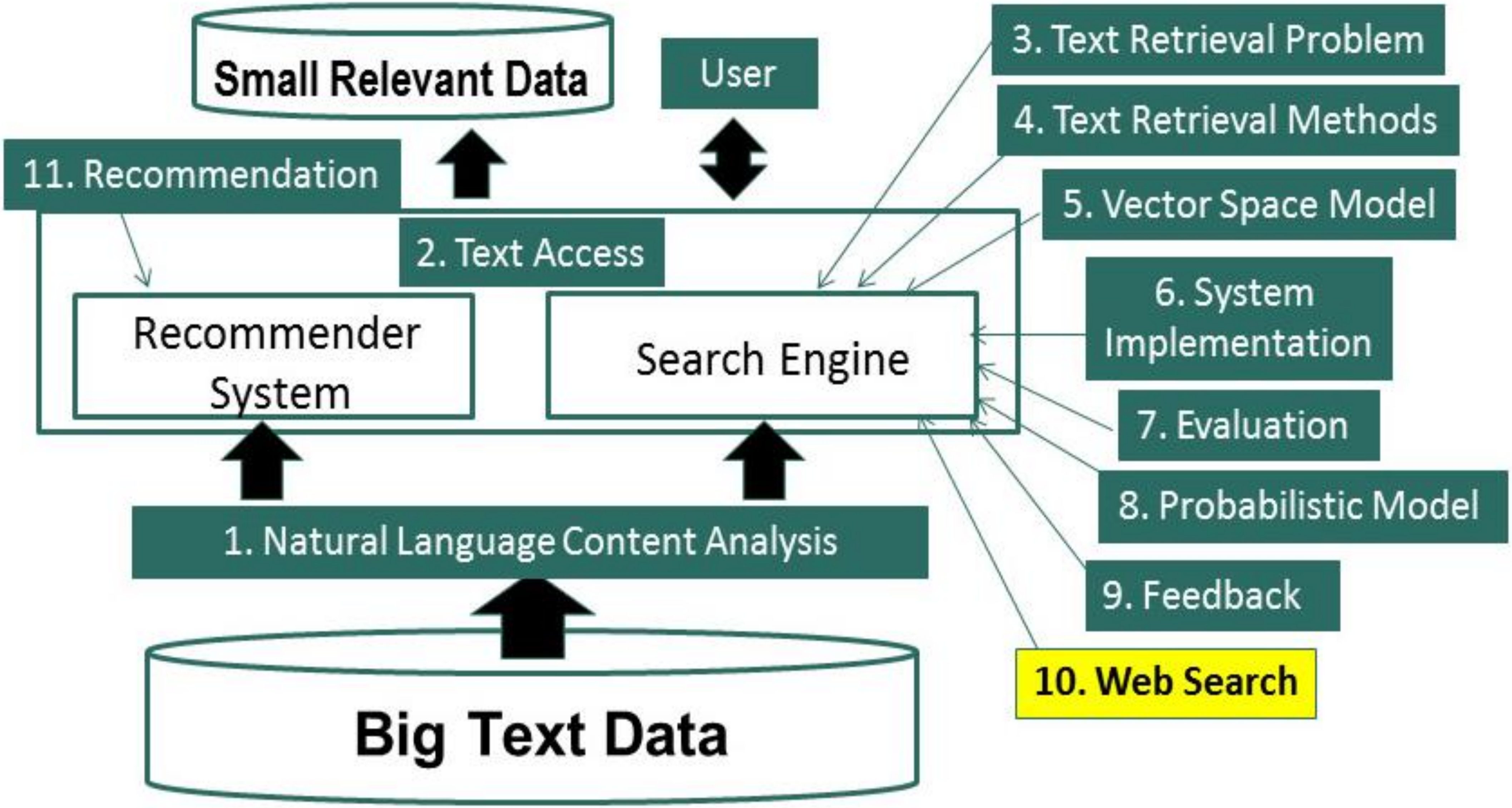
Text Retrieval and Search Engines

Web Search

ChengXiang "Cheng" Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign

Course Schedule



Web Search: Challenges & Opportunities

- Challenges
  - Scalability

    → Parallel indexing & searching (MapReduce)

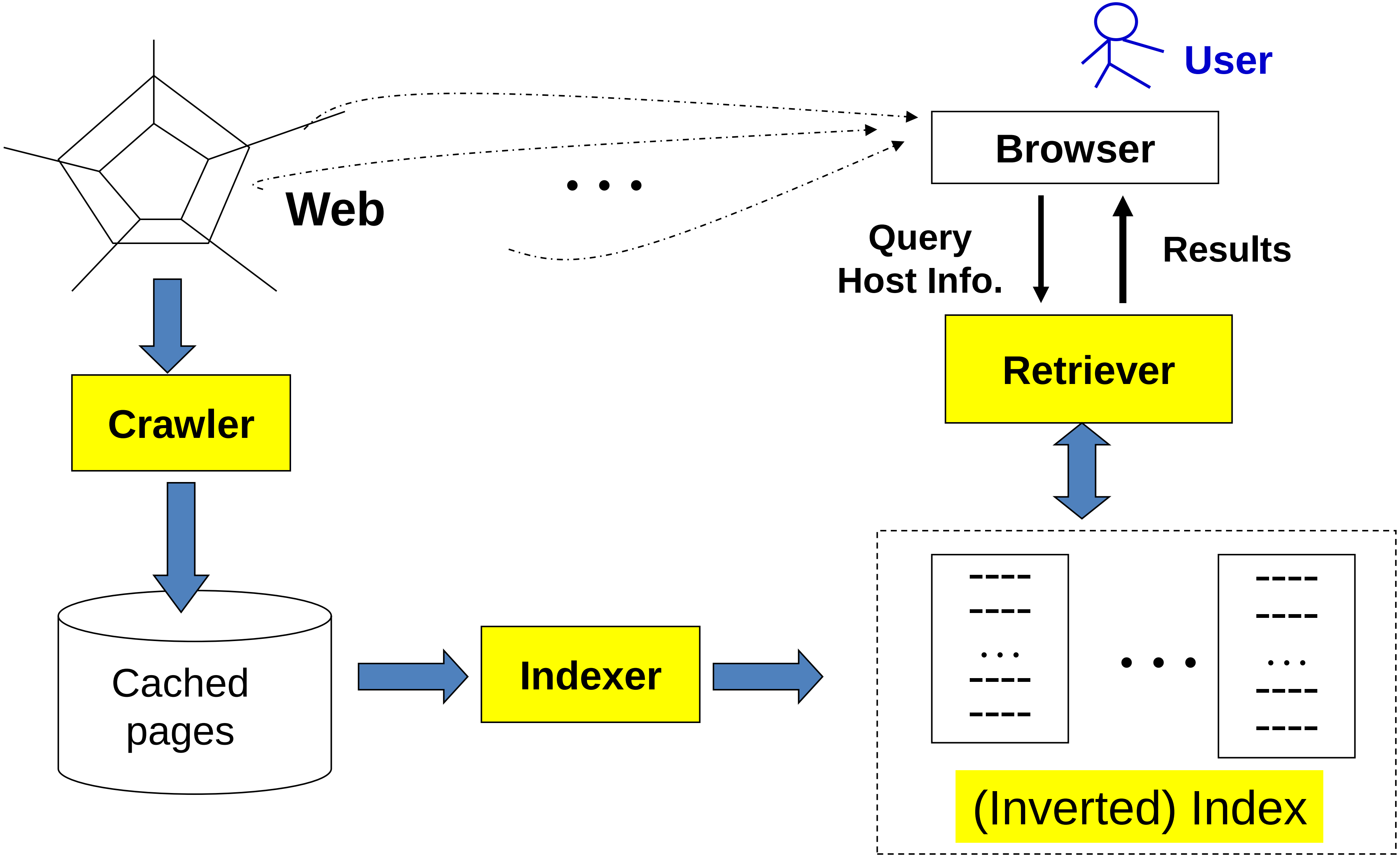
    - How to handle the size of the Web and ensure completeness of coverage?
    - How to serve many user queries quickly?
  - Low quality information and spams

    → Spam detection & Robust ranking

  - Dynamics of the Web
    - New pages are constantly created and some pages may be updated very quickly
- Opportunities
  - many additional heuristics (e.g., links) can be leveraged to improve search accuracy

    → Link analysis & multi-feature ranking

Basic Search Engine Technologies



Component I: Crawler/Spider/Robot

- Building a “toy crawler” is easy
  - Start with a set of “seed pages” in a priority queue
  - Fetch pages from the web
  - Parse fetched pages for hyperlinks; add them to the queue
  - Follow the hyperlinks in the queue
- A real crawler is much more complicated…
  - Robustness (server failure, trap, etc.)
  - Crawling courtesy (server load balance, robot exclusion, etc.)
  - Handling file types (images, PDF files, etc.)
  - URL extensions (cgi script, internal references, etc.)
  - Recognize redundant pages (identical and duplicates)
  - Discover “hidden” URLs (e.g., truncating a long URL)

Major Crawling Strategies

- Breadth-First is common (balance server load)
- Parallel crawling is natural
- Variation: focused crawling
  - Targeting at a subset of pages (e.g., all pages about “automobiles”)
  - Typically given a query
- How to find new pages (they may not linked to an old page!)
- Incremental/repeated crawling
  - Need to minimize resource overhead
  - Can learn from the past experience (updated daily vs. monthly)
  - Target at : 1) frequently updated pages; 2) frequently accessed pages

Summary

- Web search is one of the most important applications of text retrieval
  - New challenges: scalability, efficiency, quality of information
  - New opportunities: rich link information, layout, etc
- Crawler is an essential component of Web search applications
  - Initial crawling: complete vs. focused
  - Incremental crawling: resource optimization