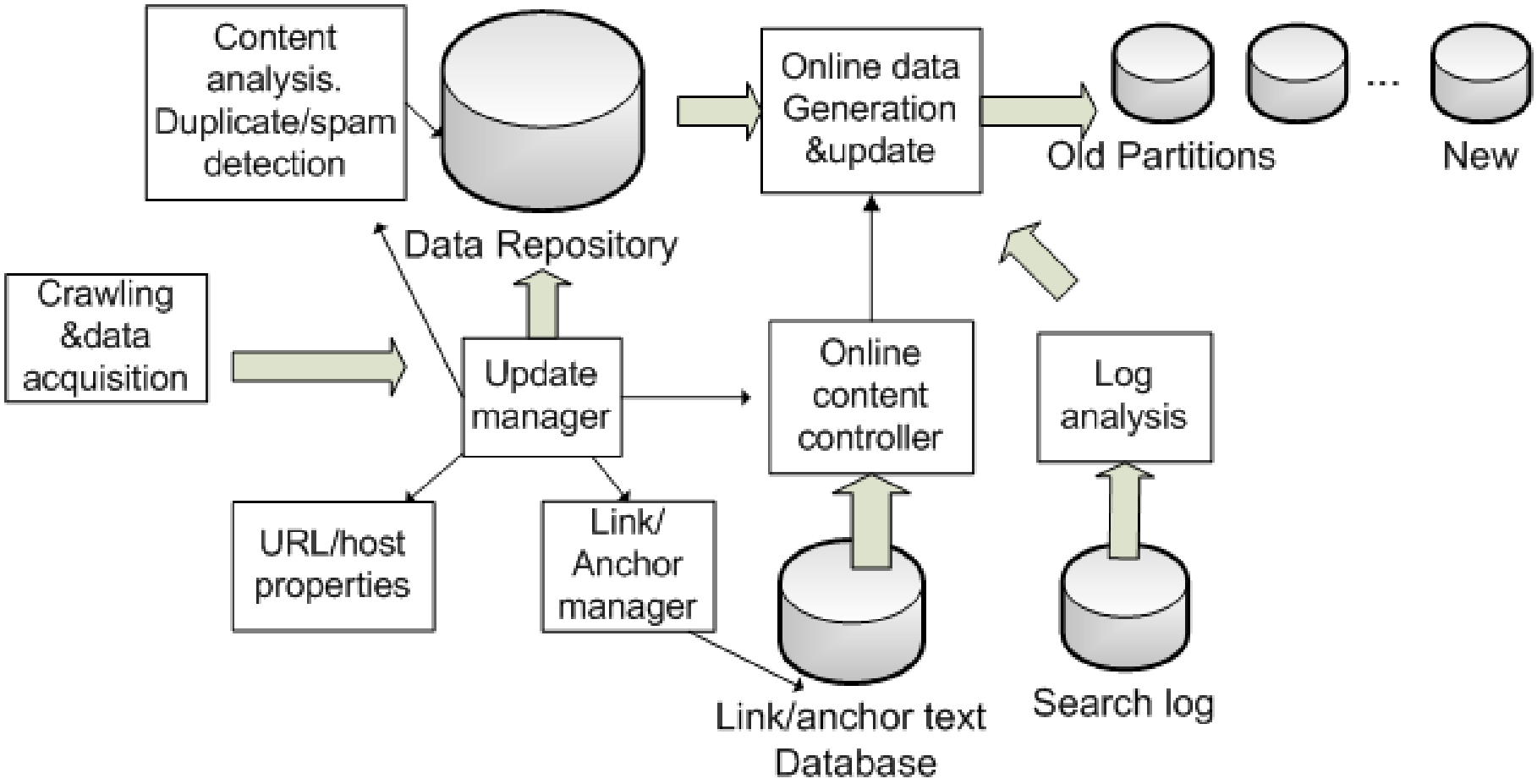
Offline Data Processing: Tasks and Infrastructure Support

T. Yang, UCSB 290N

Table of Content

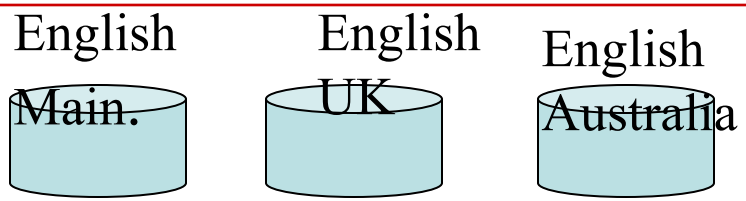
Offline Architecture for Ask.com Search



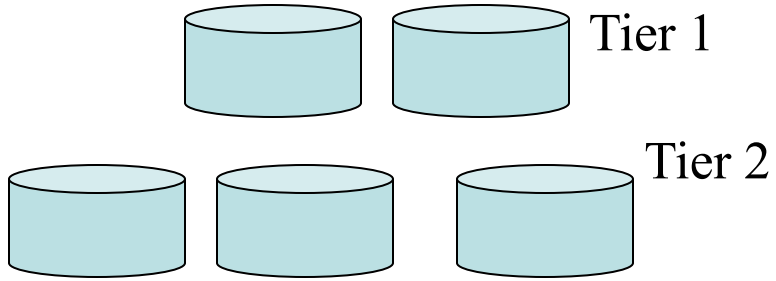
Content Management

- **Organize the vast amount of pages crawled to facilitate online search.**
  - Data preprocessing
  - Inverted index
  - Compression
  - Classify and partition data
- **Collect additional content and ranking signals.**
  - Link, anchor text, log data
- **Extract and structure content**
- **Duplicate detection**

Classifying and Partitioning data



- **Classify**
  - Content quality. Language/country etc
- **Partition**
  - Based on languages and countries. Geographical distribution based on data center locations
  - Partition based on quality
    - First tier --- high chance that users will access
      - Quality indicator
      - Click feedback
    - Second tier – lower chance

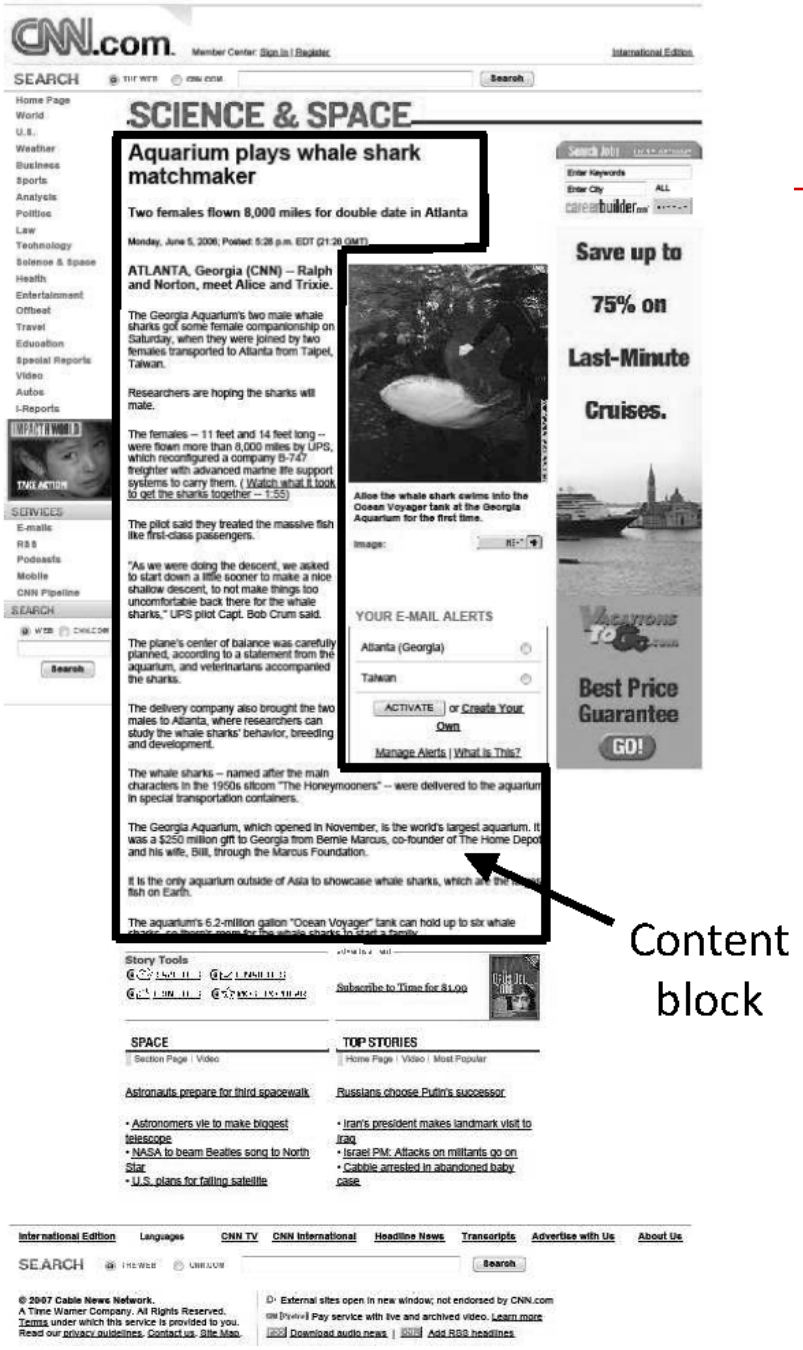


Examples of Context Extraction/Analysis

- **Identify key phrases that capture the meaning of this document.**
  - For example, title, section title, highlighted words.
    - HTML vs PDF
- **Identify parts of a document representing the meaning of this document.**
  - Many web pages contain a side-menu, which his less relevant to the main content of the documents
- **Capture page content through Javascript analysis.**
  - Page rendering and Javascript evaluation within a page

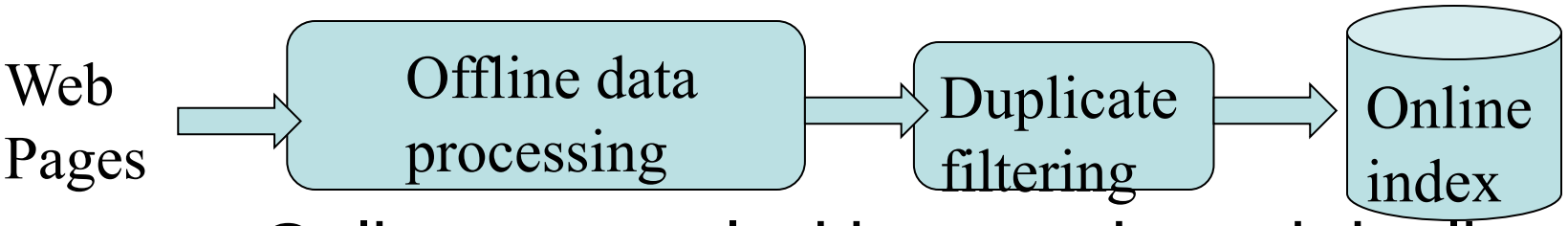
Example of Content Analysis

- **Detect content block related to the main content of a page**
  - Non-content text/link material is de-prioritized during indexing process

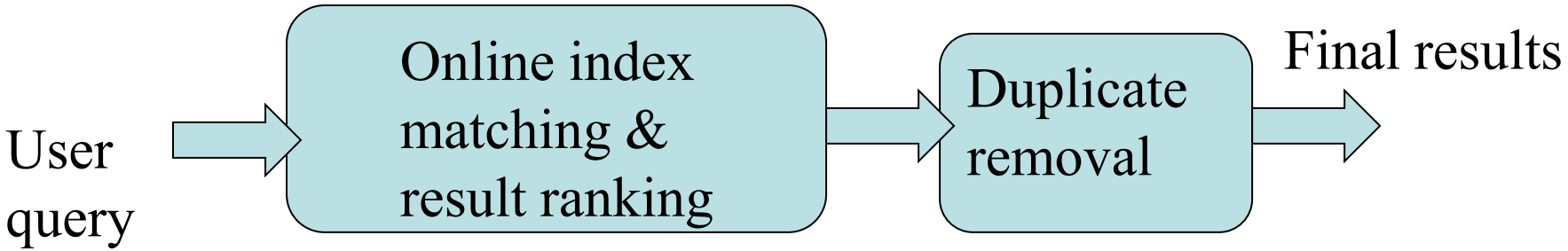


Redundant Content Removal in Search Engines

- **Over 1/3 of Web pages crawled are near duplicates**
- **When to remove near duplicates?**
  - Offline removal

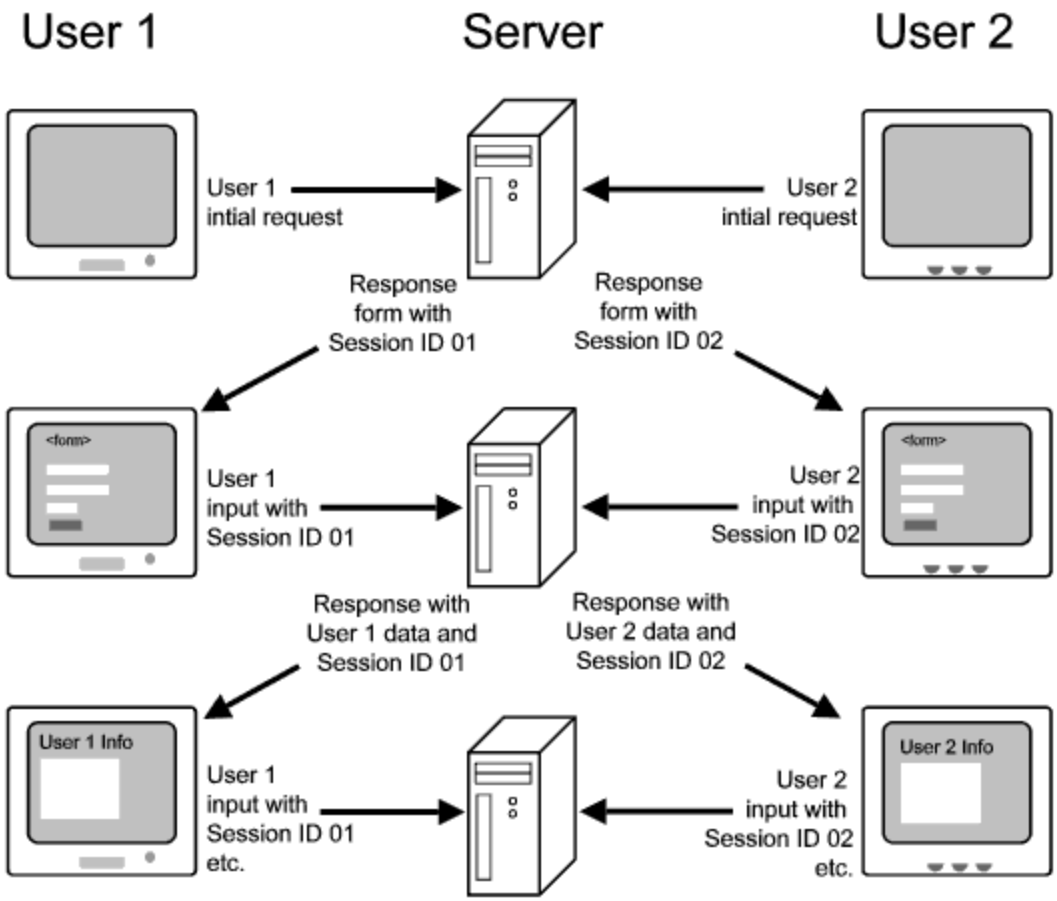


  - Online removal with query-based duplicate removal



Why there are so many duplicates?

- **Same content, different URLs, often with different session IDs.**
- **Crawling time difference**

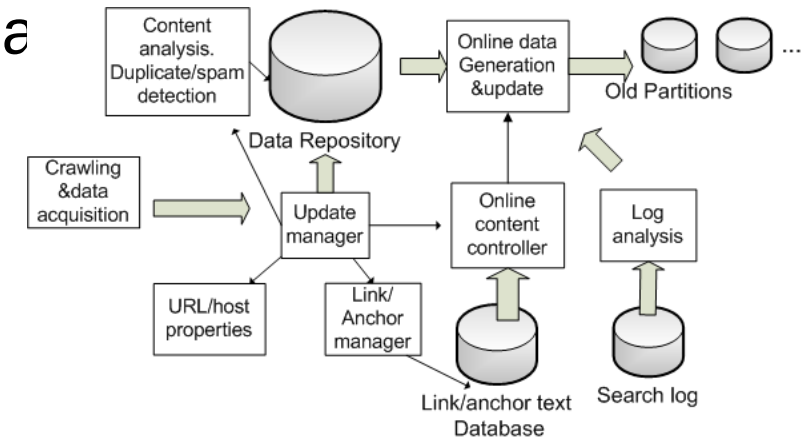


Tradeoff of online vs. offline removal

	Online-dominating approach	Offline-dominating approach
Impact to offline	High precision Low recall Remove fewer duplicates	High precision High recall Remove most of duplicates Higher offline burden
Impact to online	More burden to online deduplication	Less burden to online deduplication
Impact to overall cost	Higher serving cost	Lower serving cost

Software Infrastructure Support at Ask.com

- **Programming support (multi-threading/exception Handling, Hadoop MapReduce)**
- **Data stores for managing billions of objects**
  - Distributed hash tables, queues etc
- **Communication and data exchange among machines/services**
- **Execution environment**
  - Controllable (stop, pause, restart).
  - Service registration and invoca
  - service monitoring
  - Logging and test framework.



Requirements for Data Repository Support in Offline Systems

- **Update**
  - handling large volumes of modified documents
  - adding new content
- **Random access**
  - request the content of a document based on its URL
- **Compression and large files**
  - reducing storage requirements and efficient access
- **Scan**
  - Scan documents for text mining.

Options for Data Stores

- **Bigtable at Google**
- **Dynamo at Amazon**
- **Open source software**

	Technology	Language Platform	Users/ sponsors
Apache Cassandra	Bigtable Dynamo	Java/Hadoop	Apache
Hypertable	Bigtable	C++/Hadoop	Baidu
Hbase	Bigtable	Java/Hadoop	Apache
LevelDB	Bigtable	C++	Google
MongoDB		C++	