UNIVERSITY of DELAWARE UNIVERSITY of DE

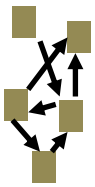
Web Search Basics

(Search and Data Mining)

Hui Fang
Department of Electrical and Computer Engineering
University of Delaware

UNIVERSITY of DELAWARE UNIVERSITY of DE

Characteristics of Web Data



- **The size of Web data is huge.**
  - Surface: static pages
  - Deep: dynamically generated pages
- **Frequent updates**
- **Variances in quality**
  - Web Spams
- **Graph-based data**

UNIVERSITY of DELAWARE UNIVERSITY of DE

General Challenges in Web Search

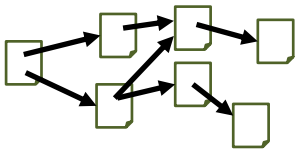
- Scalability — **Parallel query processing (MapReduce)**
  - How to handle the size of the Web?
- Addressing the dynamics of the Web — **Crawling/ Redundancy check**
  - Some pages may be updated very quickly
  - New pages are constantly created
- Dealing with low quality information — **Spam detection**
- Utilizing the additional information (such as links) to improve search quality — **Link analysis / multi-feature ranking**

3

UNIVERSITY of DELAWARE UNIVERSITY of DE

Component I: Crawler

- Basic steps of crawling
  - Start with a set of seed pages
  - Fetch pages from the Web
  - Parse the fetched pages and add the hyperlink to the list to be crawled
  - Repeat the previous steps until the list to be crawled is empty.

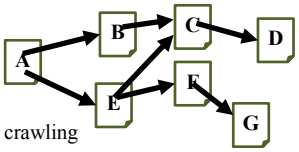


4

UNIVERSITY of DELAWARE UNIVERSITY of DE

More Details about Crawling

- Breath-First



- Parallel crawling
- Focused crawling
- Incremental/repeated crawling

5

UNIVERSITY of DELAWARE UNIVERSITY of DE

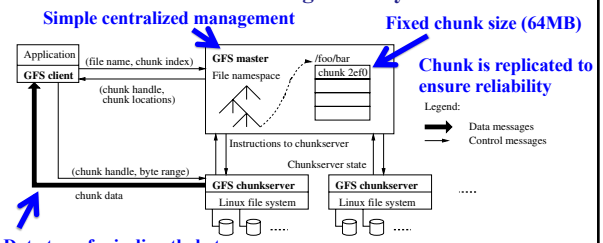
Component II: Indexing

- Standard IR techniques are the basis, but insufficient
  - Scalability
  - Efficiency
- Google's innovations on Web-scale data processing
  - Google File System
    - Storing the indexes on multiple machines
  - MapReduce
    - Creating the indexes in parallel
  - BigTable

7

UNIVERSITY of DELAWARE

Illustration of Google File System



8

UNIVERSITY of DELAWARE

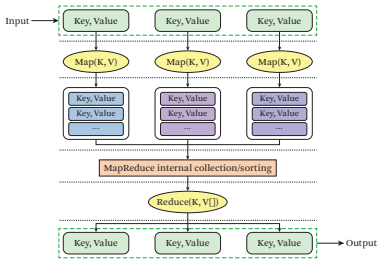
MapReduce: A Framework for Parallel Computing

- It minimizes the effort to create an application that can be run on large cluster in parallel
- Advantages
  - Hide many low-level details
  - Built-in fault tolerance

10

UNIVERSITY of DELAWARE

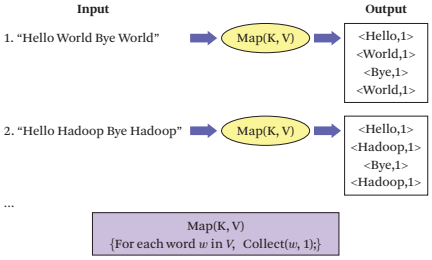
Computation Flow of MapReduce



11

UNIVERSITY of DELAWARE

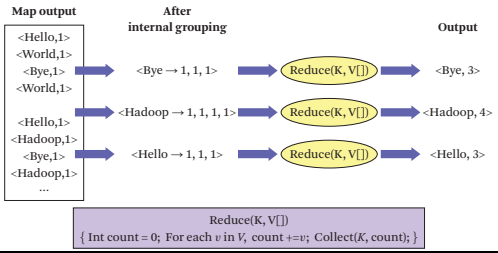
The map Function for word counting



12

UNIVERSITY of DELAWARE

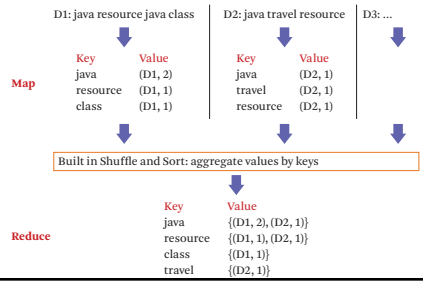
The reduce Function for word counting



13

UNIVERSITY of DELAWARE

Using MapReduce to create an inverted index



14

Pseudocode for Inverted Index Construction

```
function MAP(docid n, doc d)
    H ← new ASSOCIATIVEARRAY
    for term t ∈ d do
        H[t] ← H[t] + 1
    end for
    for t ∈ H do
        EMIT(t, [n, H[t]])
    end for
end function

function REDUCE(term t, postings [(a1, f1), (a2, f2), ...])
    P ← new LIST
    for all (a, f) do
        APPEND(P, (a, f))
    end for
    SORT(P)
    EMIT(t, P)
end function
```

15

Component III: Retriever

- Traditional IR models are not enough
  - Information needs are different: navigational vs. informational
  - Documents have additional information
  - Information quality varies a lot
- Specific techniques for Web search
  - Exploiting links
  - Spelling correction
  - Spam detection
  - Leveraging more features such as clickthrough data
- In general, rely on machine learning to combine all kinds of features.

17