Web Data Management

Introduction

Web Data Management(WDM)

- A body of work concerned with leveraging the large collections of structured data that can be extracted from the Web

- Exploring these collections of data with the goal of improvingWeb search and developing mechanisms for surfacing different kinds of search answers.

Data Type

- Text
- Video
- Audio
- Image
- Web content data
- Web link data
- Web log data

Five Topics

- Domain-Independent Extraction fromText
- Online Data Communities
- Social Data Management Tools
- The DeepWeb
- Web Search and Information Retrieval

Domain-Independent Extraction from Text

- Extraction systems that are able to effectively construct relational databases out of very large document corpora

- A large number of recent and relevant academic projects

Projects

- WebTables system
  - A large corpus of databases from HTML tables on the Web.
- IIT Bombay
  - annotate tabular data elements with extra semantic information (e.g., the type of a column)

Online Data Communities

- Socially-driven data creation systems
- Wikipedia data has become a critical standard in most socially-driven data work

Online Data Communities

- Freebase is a community-constructed graph-oriented database
- DBPedia is an effort to unify several online structured databases
- MusicBrainz for music, Geonames for geographic information, Drugbank for pharmaceutical information

Social Data Management Tools

- FusionTables is a Google tool that enables socially-driven creation of tabular datasets
- IBM's ManyEyes site allows groups of people to discuss and visualize data sets.
- Socrata offers tools for mashing up and visualizing uploaded datasets, in particular governmental data.
- The DBLife system allows groups of people to easily design a topic-specific website that collects much of its data from external sources

The DeepWeb

- The Deep Web is the collection of databases with Web front-ends, containing data that can only be accessed via submitted Web forms.

- Many estimates of the Deep Web put its size at several times the data that can be accessed via the traditional Web.

Related Work

- Data mining is occupied with techniques for obtaining high-quality predictive or other statistical results from examining datasets

- Information extraction focuses on obtaining a refined version of data from an unstructured source

Web Data Exercise 1

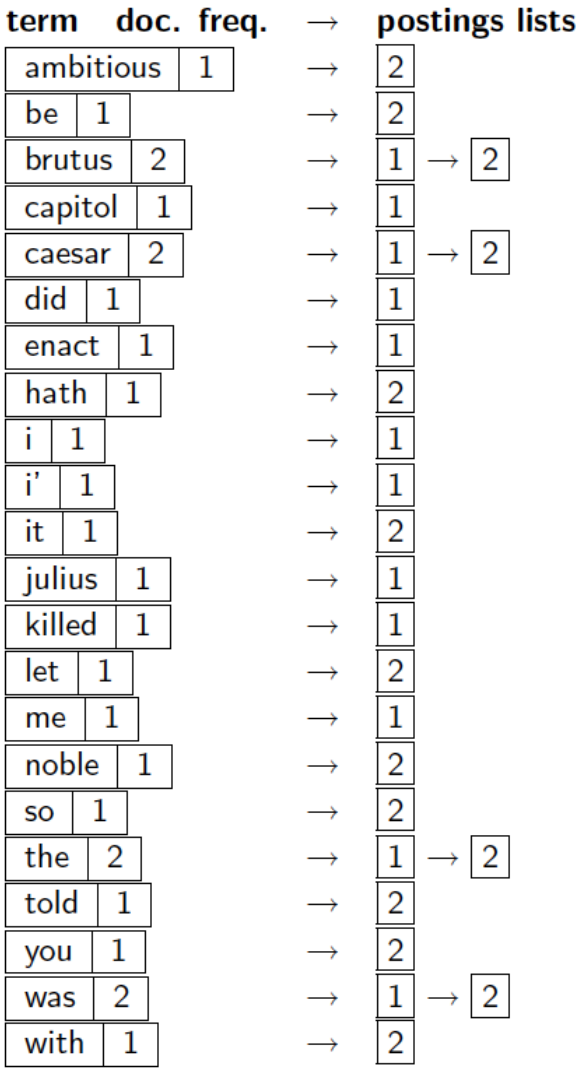
Consider the following documents:

- d1 = I like to watch the sun set with my friend.
- d2 = The Best Places To Watch The Sunset.
- d3 = My friend watch the sun come up.

Write a program which can output the document IDs given an input keyword.

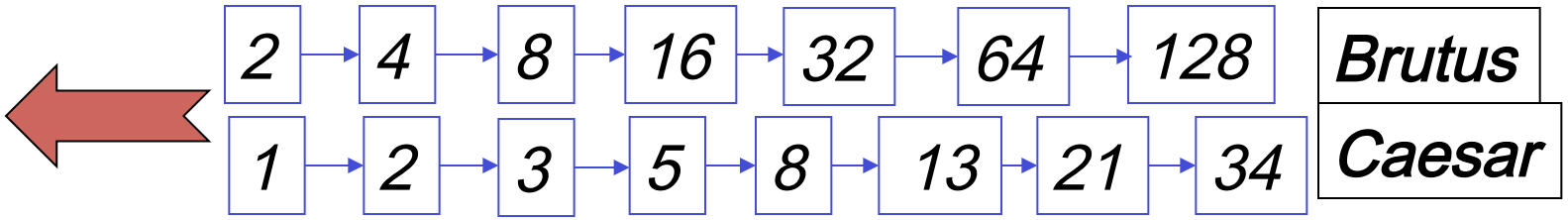
Web Data Exercise 2

- Data structure for inverted index
- Input : documents, keywords
- Output: document IDs
- HashMap



Web Data Exercise 3

- Answer X AND Y query in O(x+y) operations



- Answer X AND Y NOT Z in linear time
- Use Jaccard for ranking

Intersecting two postings lists (a "merge" algorithm)

```
INTERSECT(p1, p2)
 1  answer ← ⟨ ⟩
 2  while p1 ≠ NIL and p2 ≠ NIL
 3  do if docID(p1) = docID(p2)
 4        then ADD(answer, docID(p1))
 5             p1 ← next(p1)
 6             p2 ← next(p2)
 7        else if docID(p1) < docID(p2)
 8                then p1 ← next(p1)
 9                else p2 ← next(p2)
10  return answer
```

Web Data Exercise 4

Consider the following documents:

- d1 = I like to watch the sun set with my friend.
- d2 = The Best Places To Watch The Sunset.
- d3 = My friend watch the sun come up.

Write a program which can output the ranked list of document IDs given an input keyword.

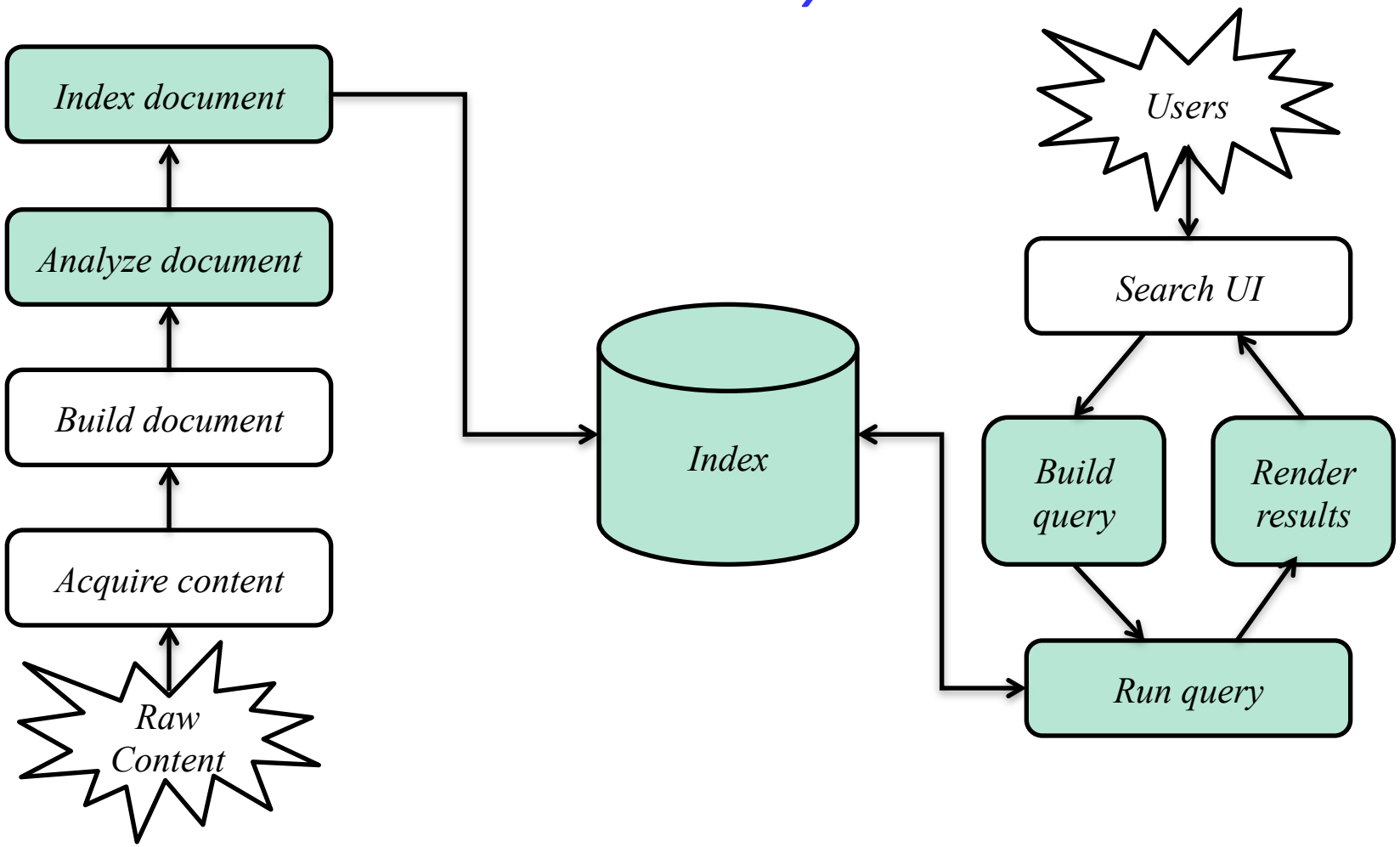
Web Data Exercise 5

- Design a crawler that can download Web pages following hyperlinks automatically
  - Input: a seed web page
  - Output: URLs from its hyperlinks
- Design a html parser that can extract titles from a Web page
  - Input: URLs from hyperlinks
  - Output: title in each URL

http://english.whut.edu.cn

Web Data Exercise 6

Lucene in a search system



cathylilin@whut.edu.cn