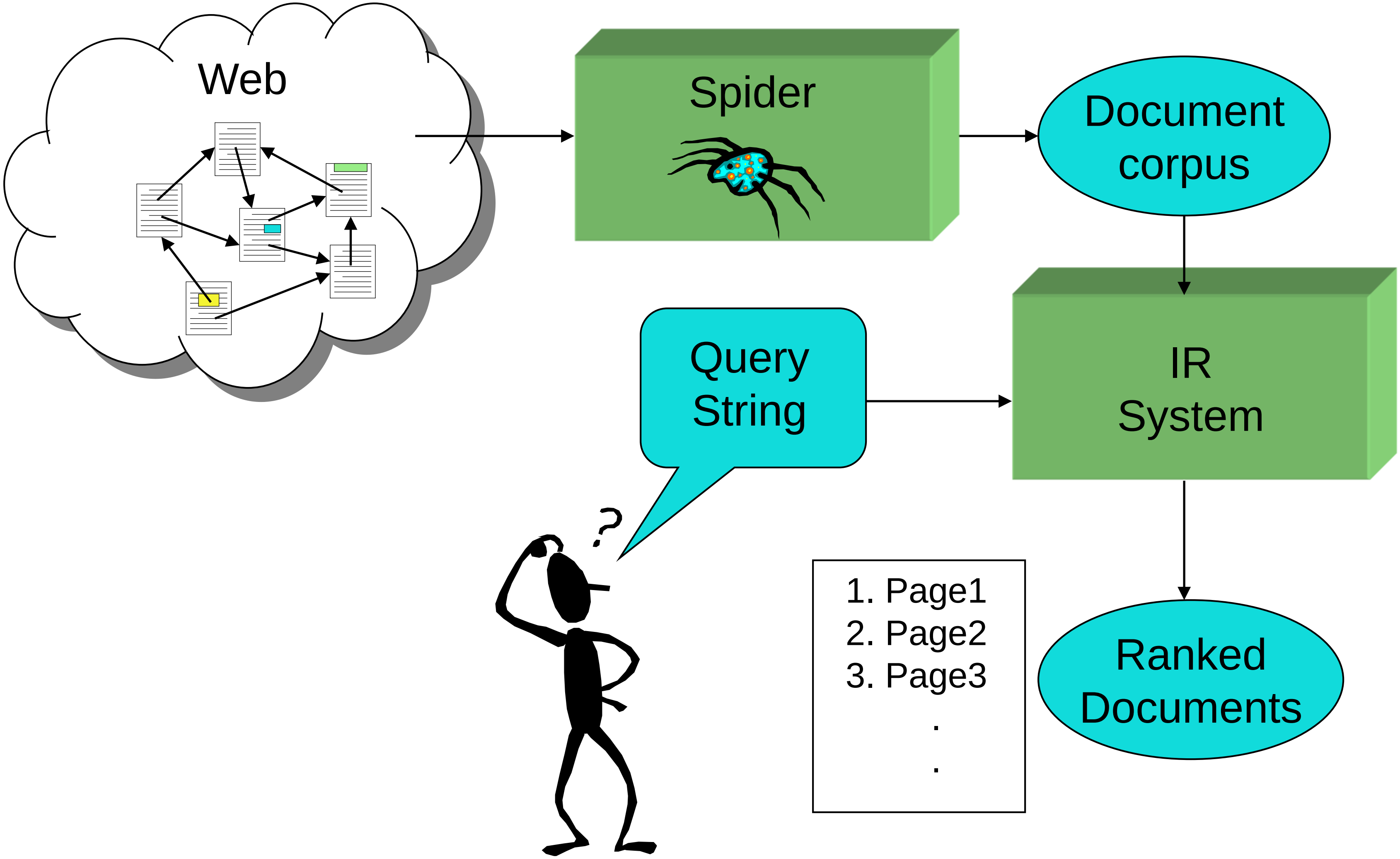
Chapter 9
Searching The Web

Web Search Using IR



Web

Spider

Document corpus

Query String

IR System

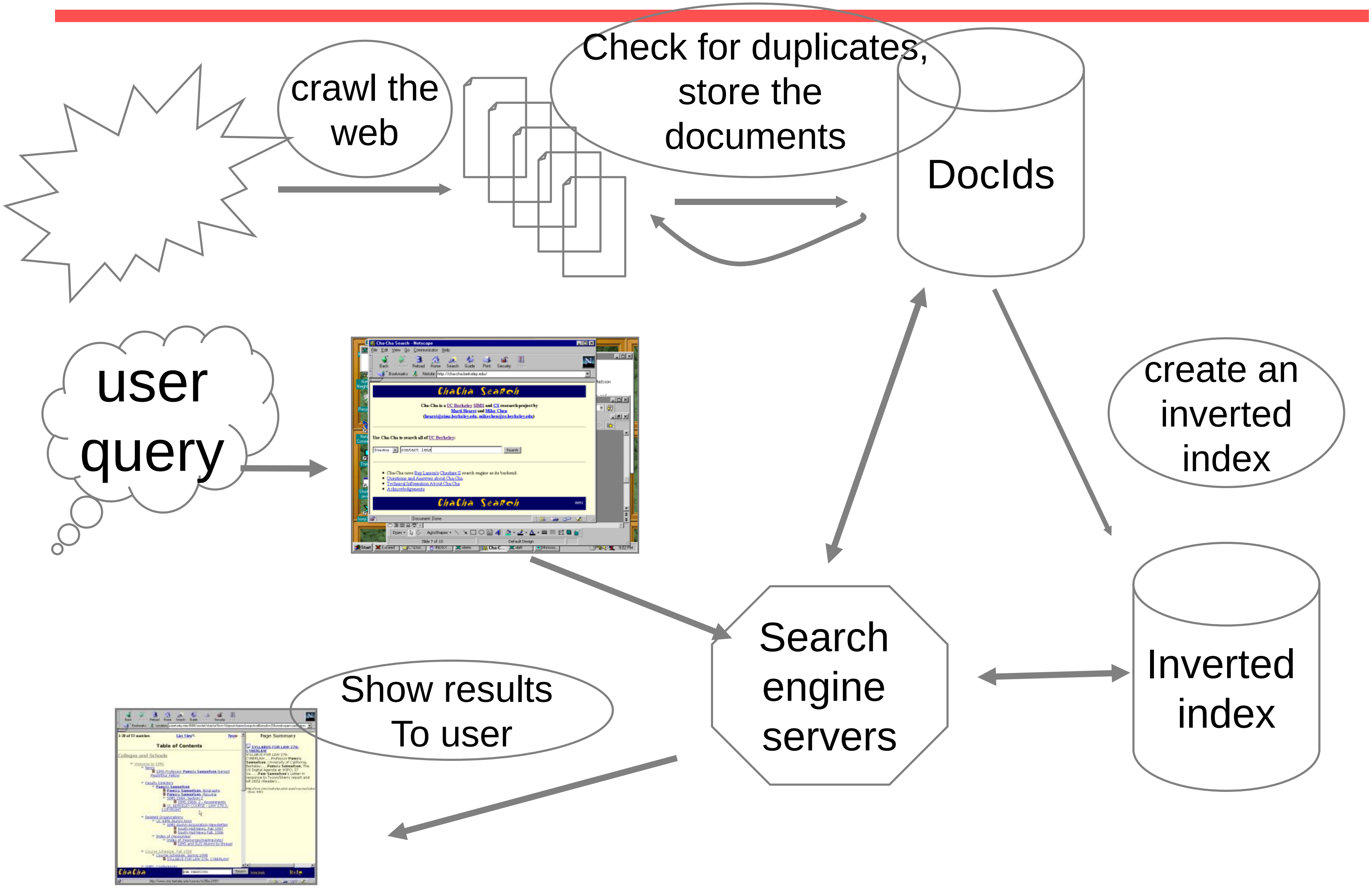
1. Page1
2. Page2
3. Page3

.

.

Ranked Documents

Standard Web Search Engine Architecture



Cha-Cha Search - Netscape

ChaCha Search

Cha-Cha is a UC Berkeley SIMS and CS research project by Marti Hearst and Mike Chen (hearst@sims.berkeley.edu, mikechen@cs.berkeley.edu)

Use Cha-Cha to search all of UC Berkeley:

Directory | contact lens | Search

- Cha-Cha uses Ray Larson's Cheshire II search engine as its backend.
- Questions and Answers about Cha-Cha
- Technical Information About Cha-Cha
- Acknowledgements

1-20 of 33 matches

Table of Contents

Colleges and Schools

Page Summary

Challenges

1. Distributed Data
2. High percentage of volatile data
3. Large volume
4. Unstructure and redundant data
5. Heterogeneous data

different languages

Search Engines

1. Centralized Architecture
2. Distributed Architecture

Centralized Architecture
(Crawler-indexer)

Definition

1. Crawlers are program (software agents) that traverse the Web sending new or updated pages to a main server where they are indexed.

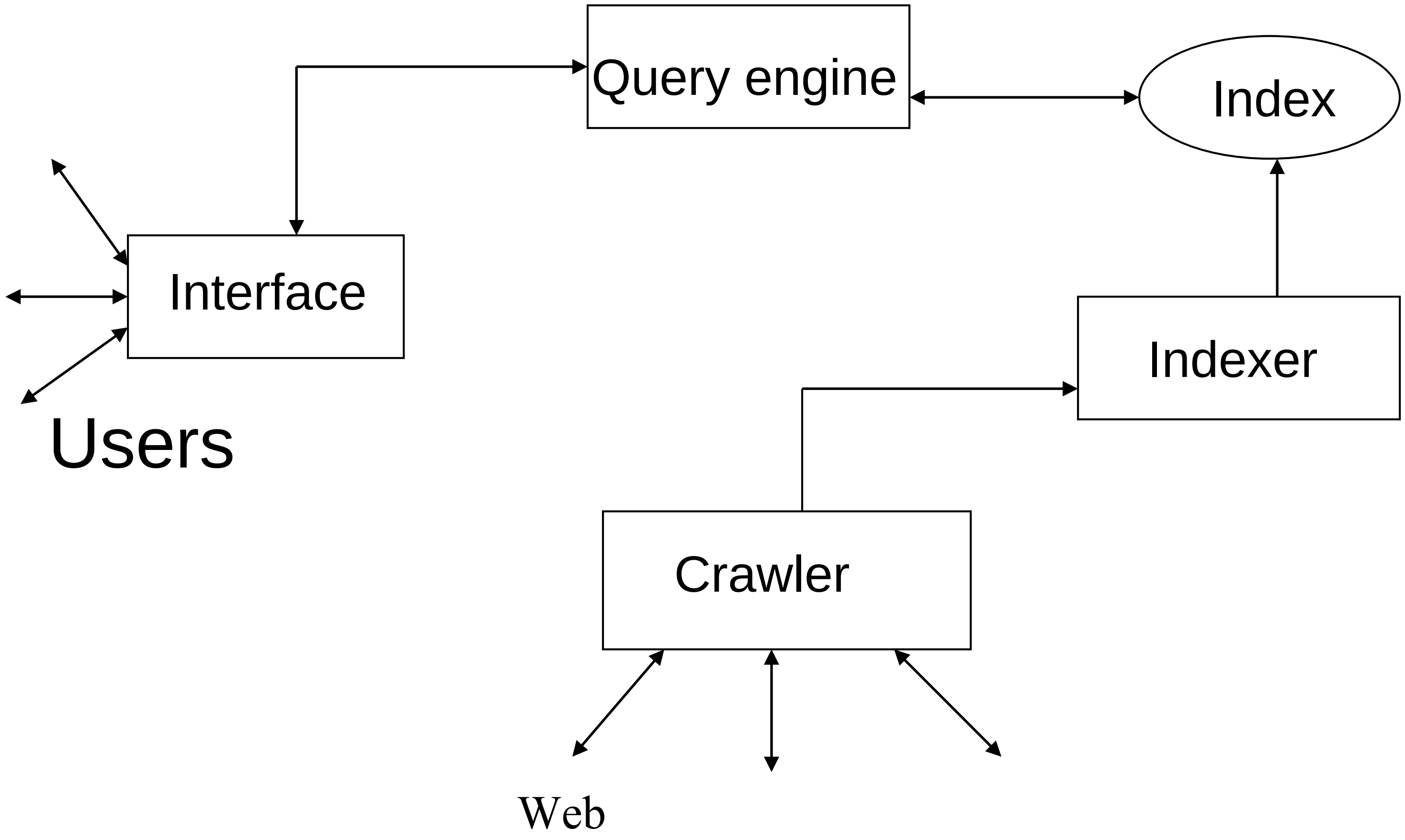
2. Run on local server and send request to remote servers

3. Centralised use of index to answer queries

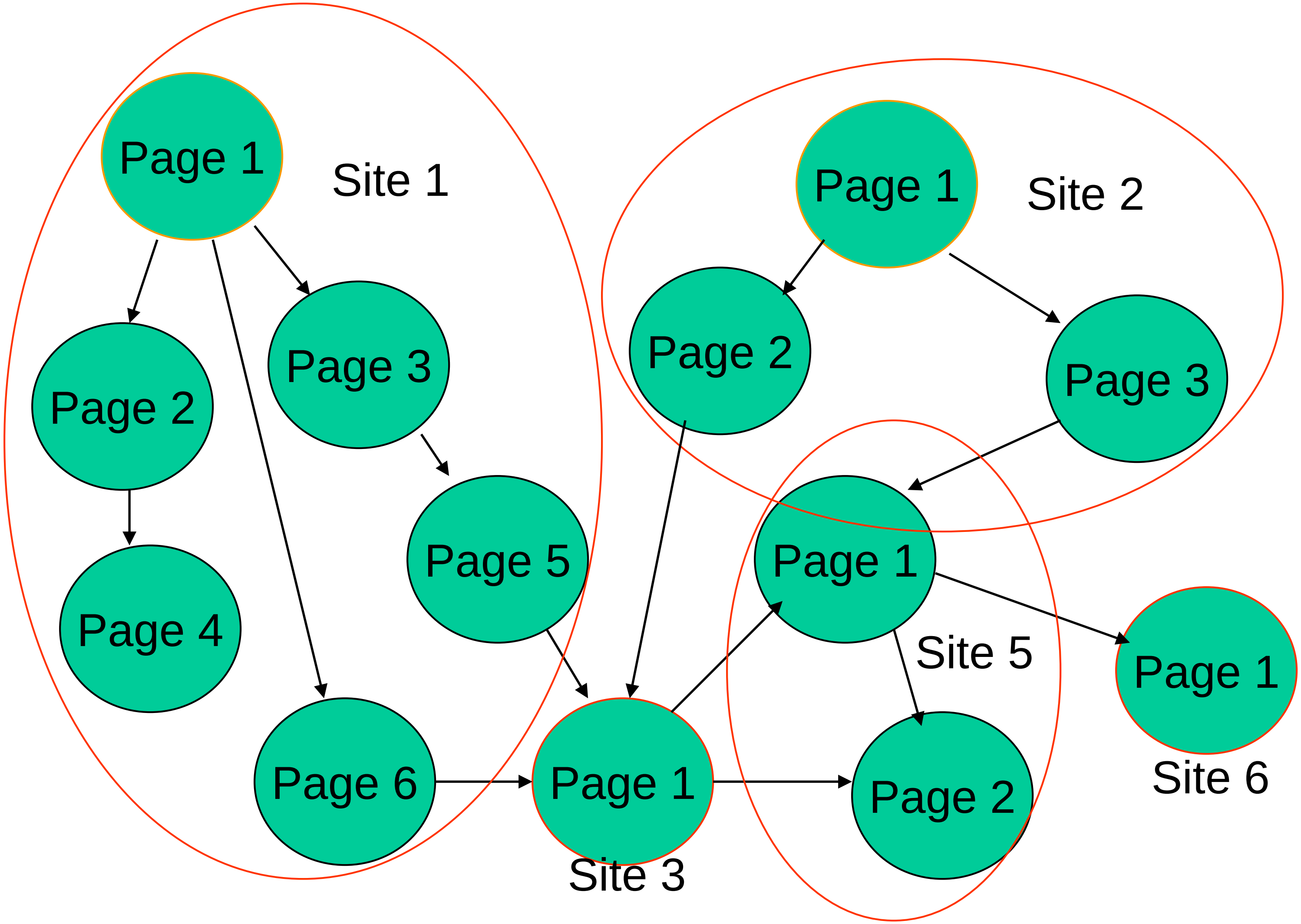
Name

Robots, Spiders, Wanderers, Walkers , Knowbot

Centralized Architecture (Crawler-indexer)

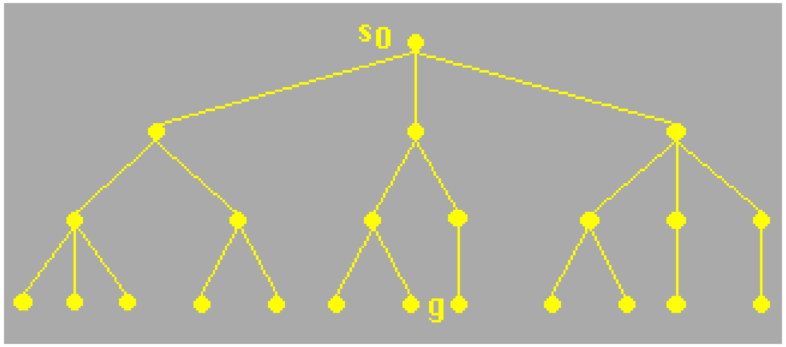


Depth-First Crawling (more complex – graphs & sites)

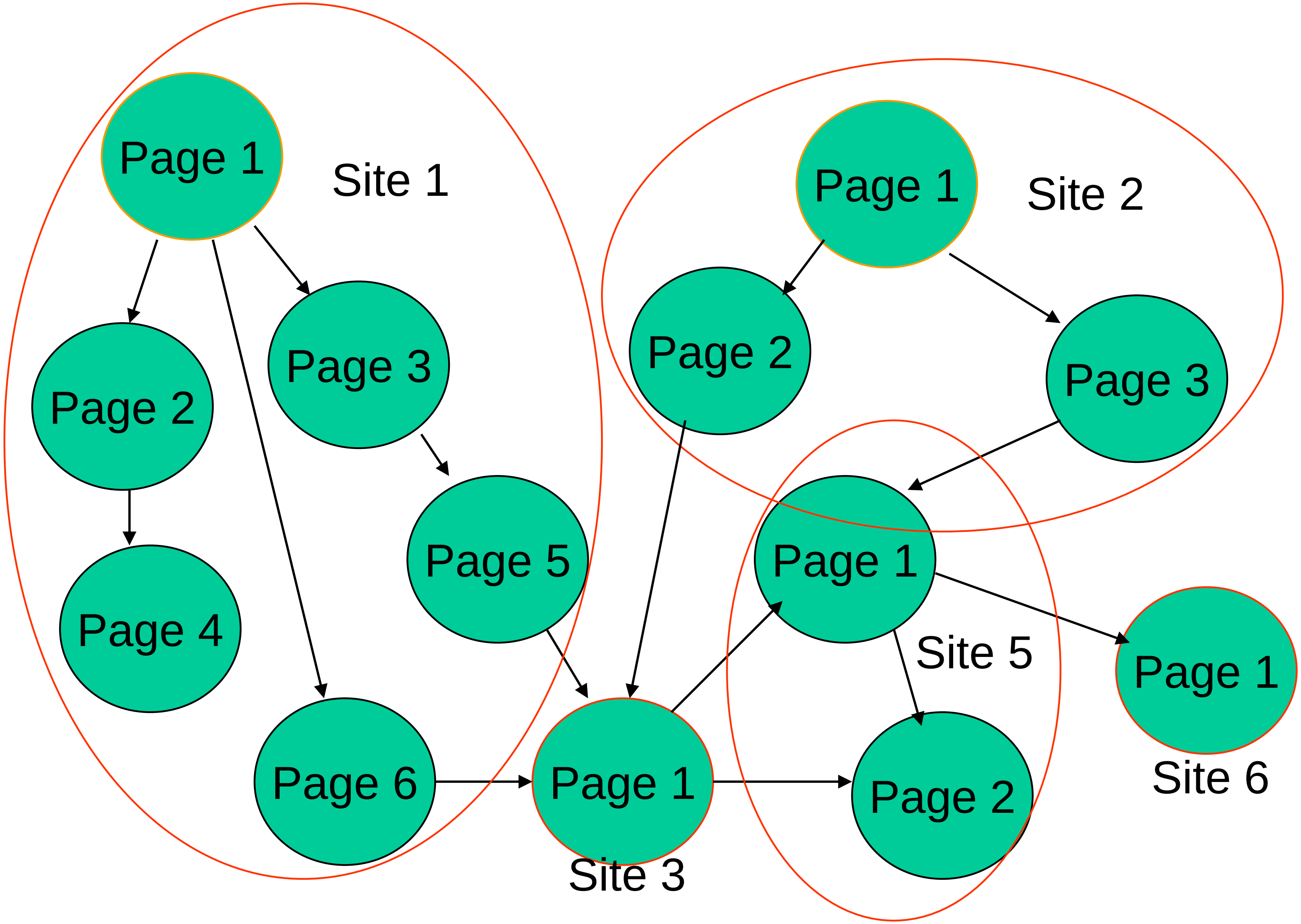


Site	Page
1	1
1	2
1	4
1	6
1	3
1	5
3	1
5	1
6	1
5	2
2	1
2	2
2	3

Depth-first search

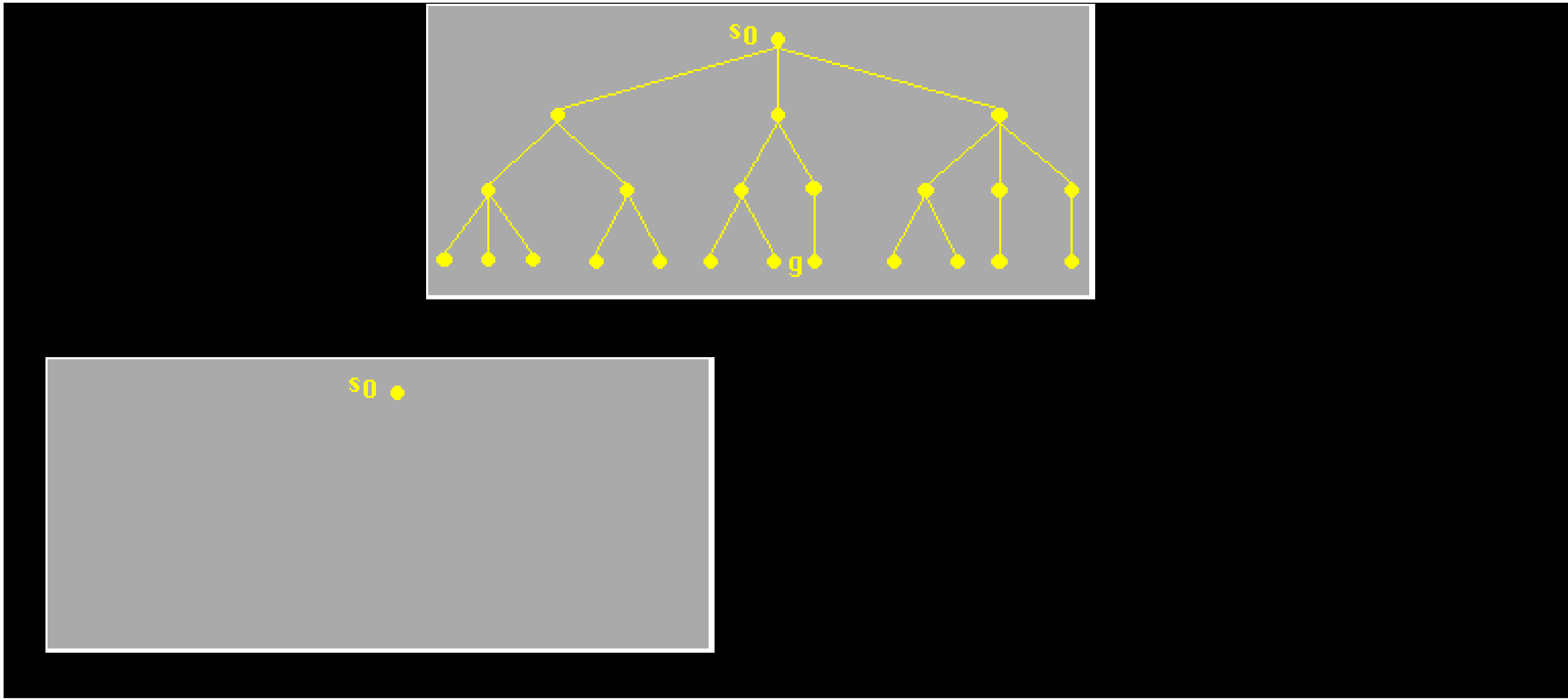


Breadth First Crawling (more complex – graphs & sites)



Site	Page
1	1
2	1
1	2
1	6
1	3
2	2
2	3
1	4
3	1
1	5
5	1
5	2
6	1

Breadth-first search



Centralized Architecture
(Crawler-indexer)

Problem

1. Volumn of the data
2. Traffic (Crawler retrieve entire object)
3. High load at Web Servers

Distributed Architecture (Harvest)

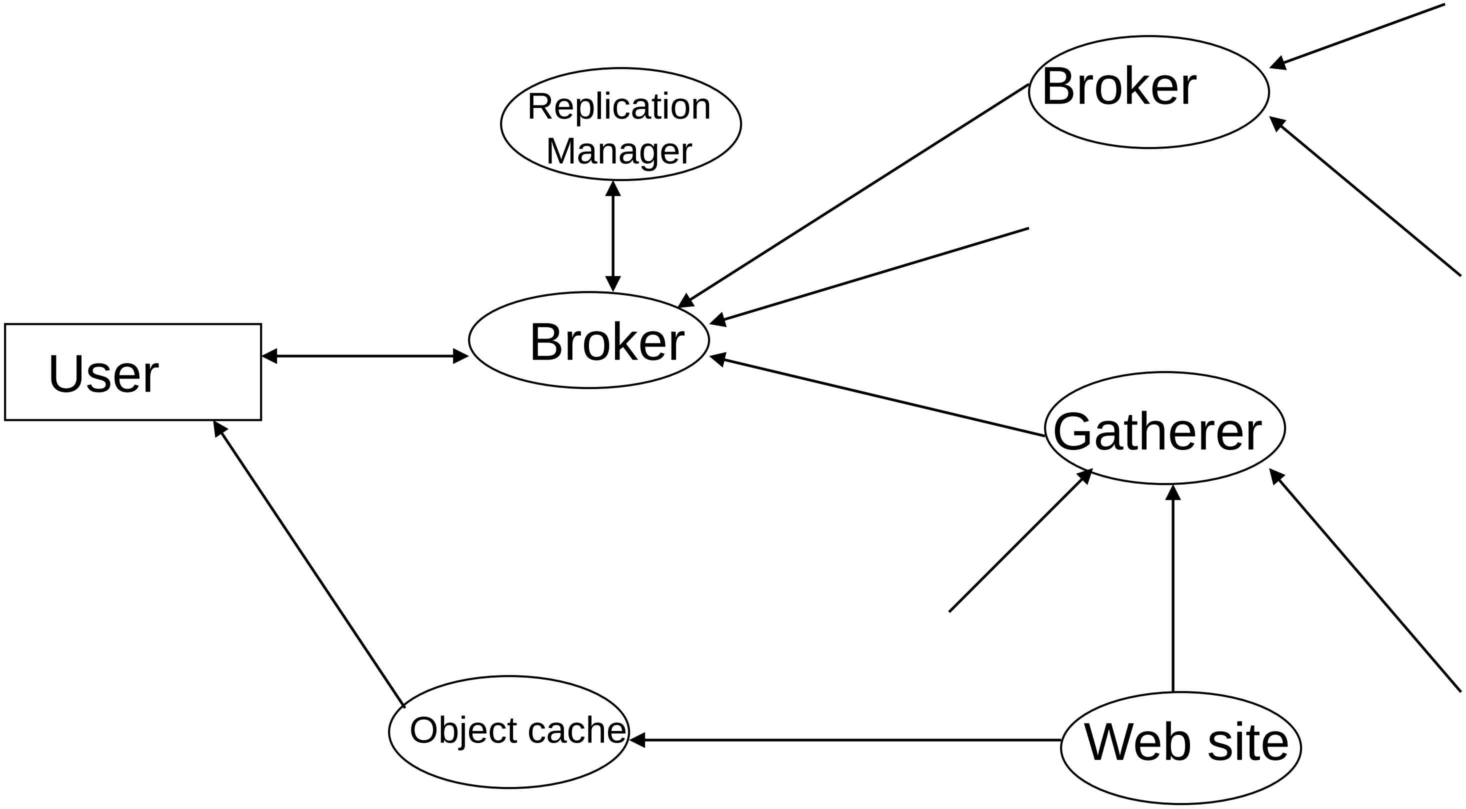
Definition

1. Gatherers collects and extracts indexing information form one or more Web servers at periodic time
2. Brokers

   -Provide indexing mechanism and query interface to data gathered

   -Retrieve information from gatherers or other brokers, updating incrementally their indices

Distributed Architecture
(Harvest architecture)


Replication Manager
Broker
User
Broker
Gatherer
Object cache
Web site