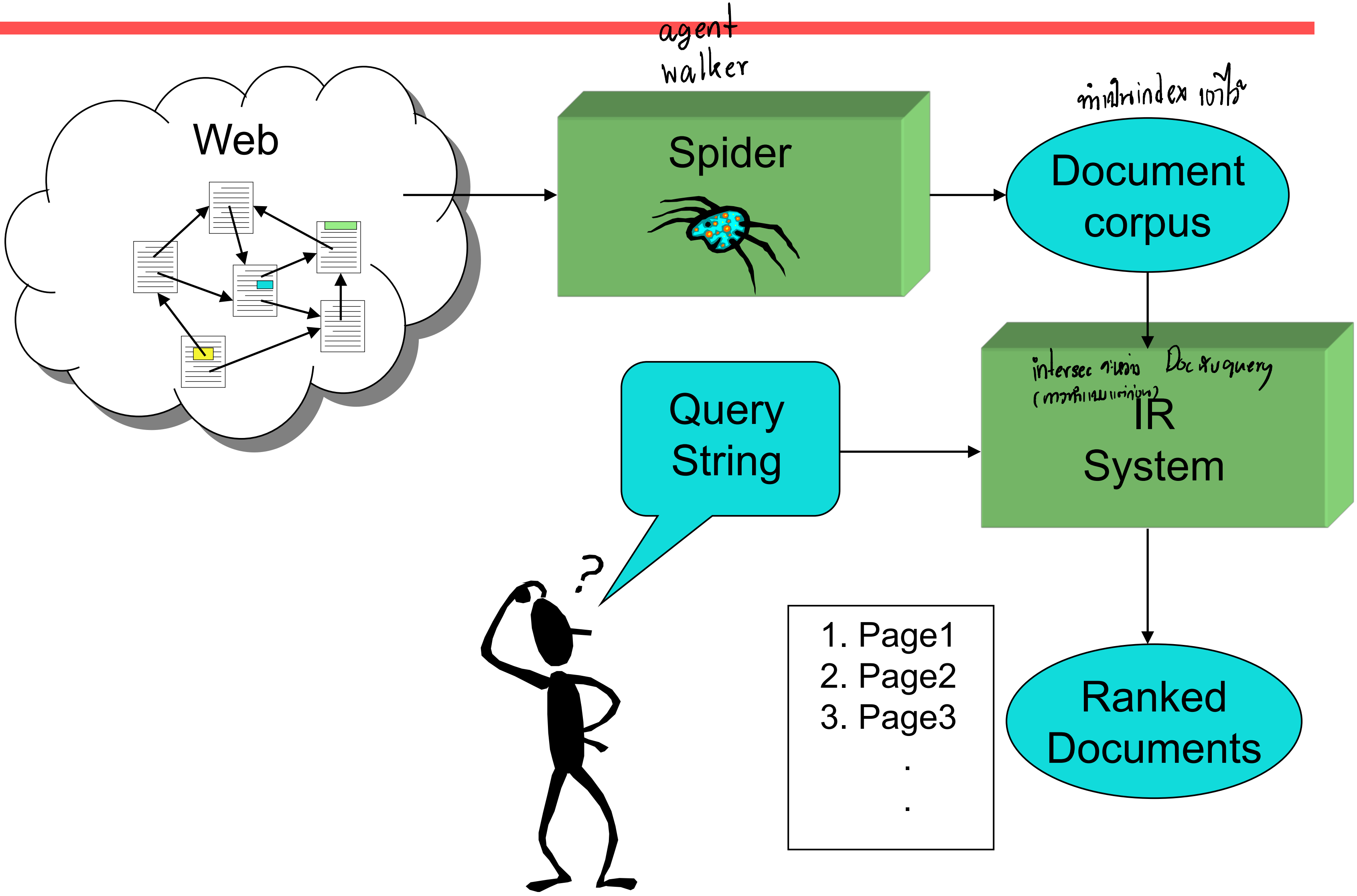
Au 1

Chapter 9
Searching The Web

Aug 3:35

Web Search Using IR



agent
walker

Web

Spider

ทำไฟล์index เอาไว้

Document corpus

intersec ระหว่าง Doc กับ query
(การทำแบบแต่ก่อน)

IR System

Query String

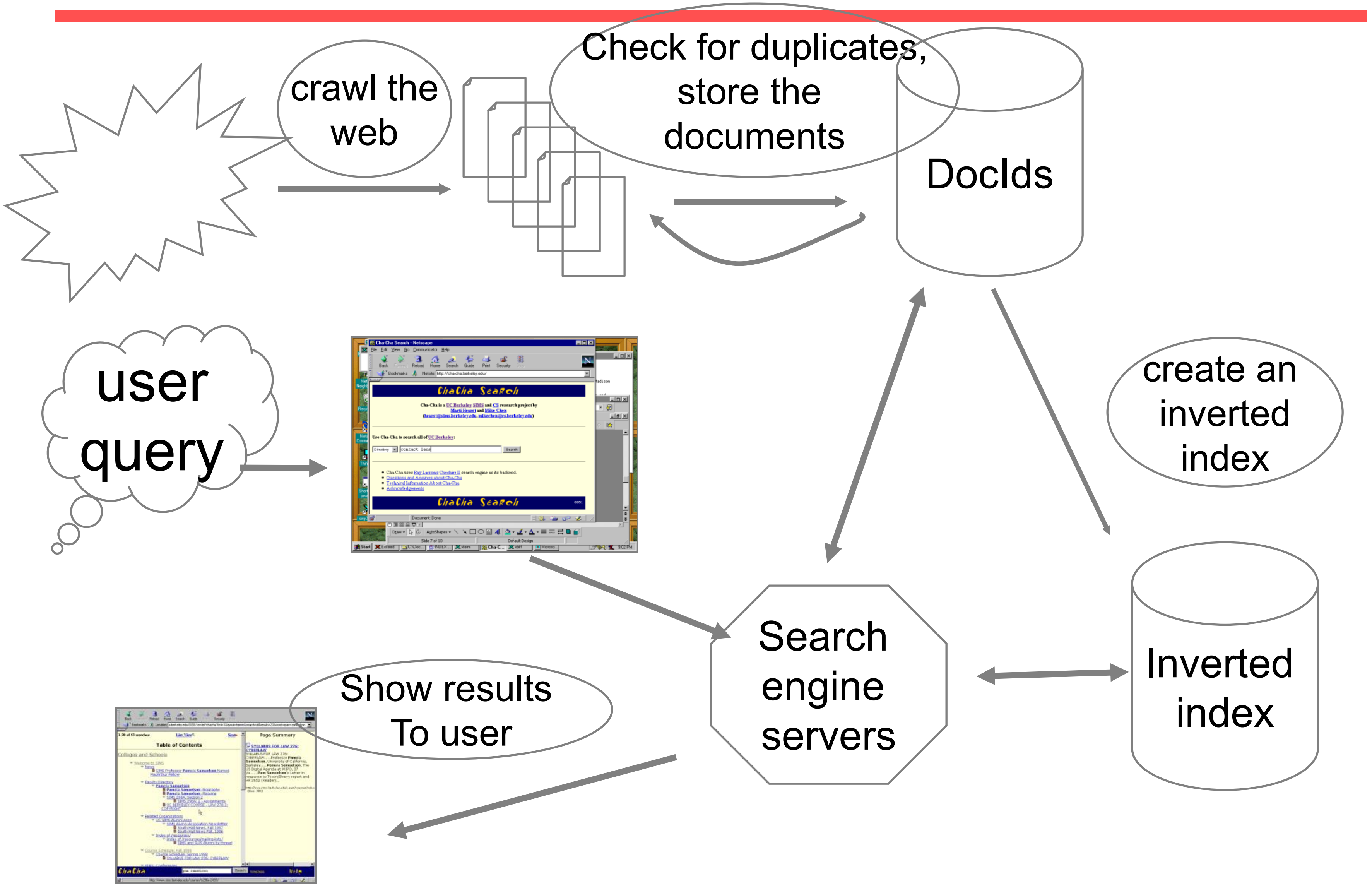
?

1. Page1
2. Page2
3. Page3
.
.

Ranked Documents

4:24

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

5:44

Challenges

ความท้าทาย ที่อยู่ในการค้นหาบนเว็บ

1. Distributed Data ข้อมูลกระจายหลาย node (หน้าที่ของ search ต้องหาให้ user ครบถ้วน)
2. High percentage of volatile data เปลี่ยนตลอดเวลา (หน้าที่ ...ต้องตอบสนอง user ให้ได้)
3. Large volume
4. Unstructure and redundant data บางอันมีโครงสร้าง / บางอันไม่มีโครงสร้าง มันซ้ำซ้อน
5. Heterogeneous data

เก็บด้วยภาษาต่างกัน

different languages

12:00 ส

Search Engines

ทำ index ไว้ตรงกลาง

1. Centralized Architecture
2. Distributed Architecture กระจายออกไป

13:20

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.

รันที่ ศูนย์กลาง

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

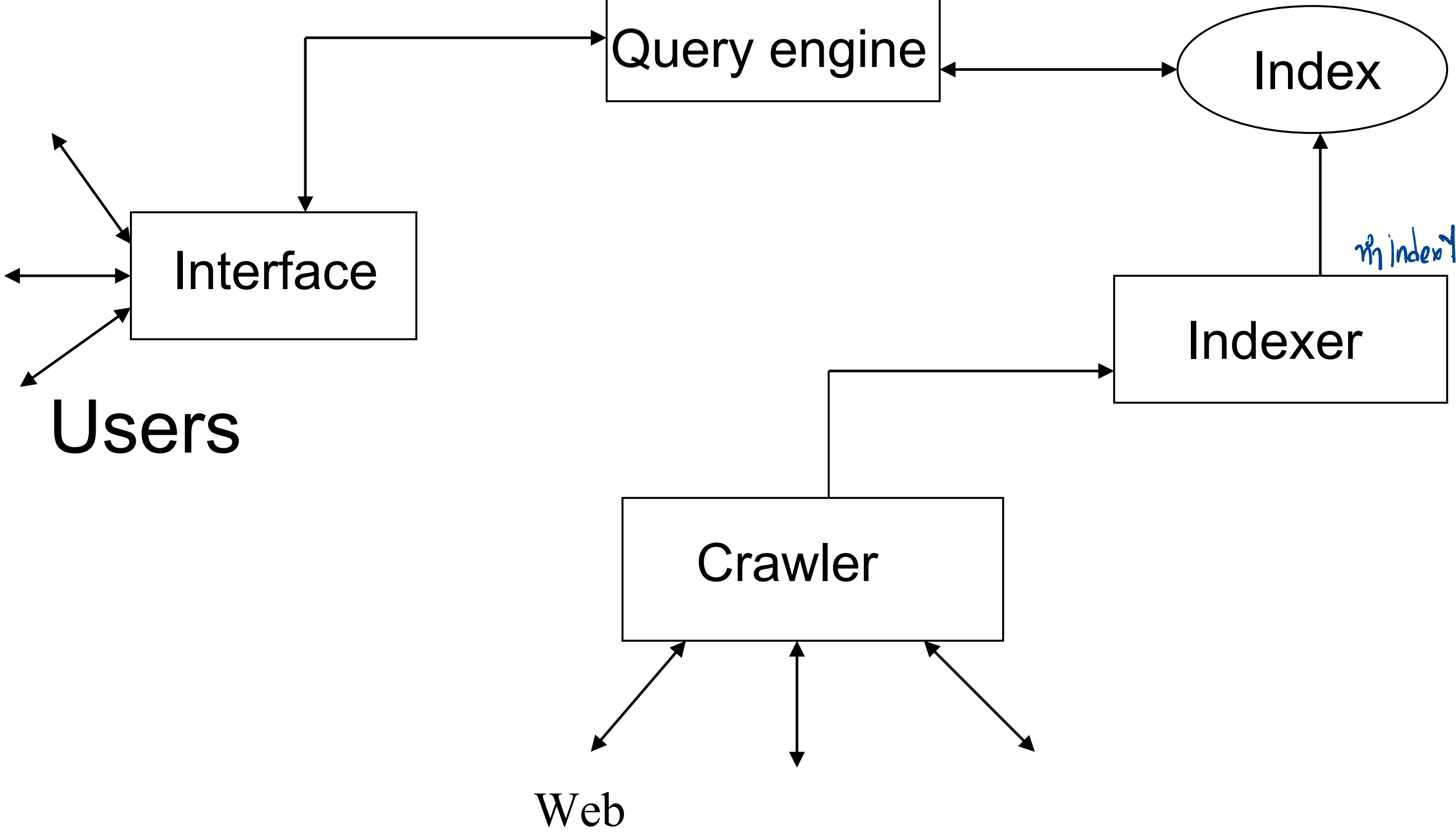
3. Centralised use of index to answer queries

Name

Robots, Spiders, Wanderers, Walkers , Knowbot

17:47

Centralized Architecture (Crawler-indexer)



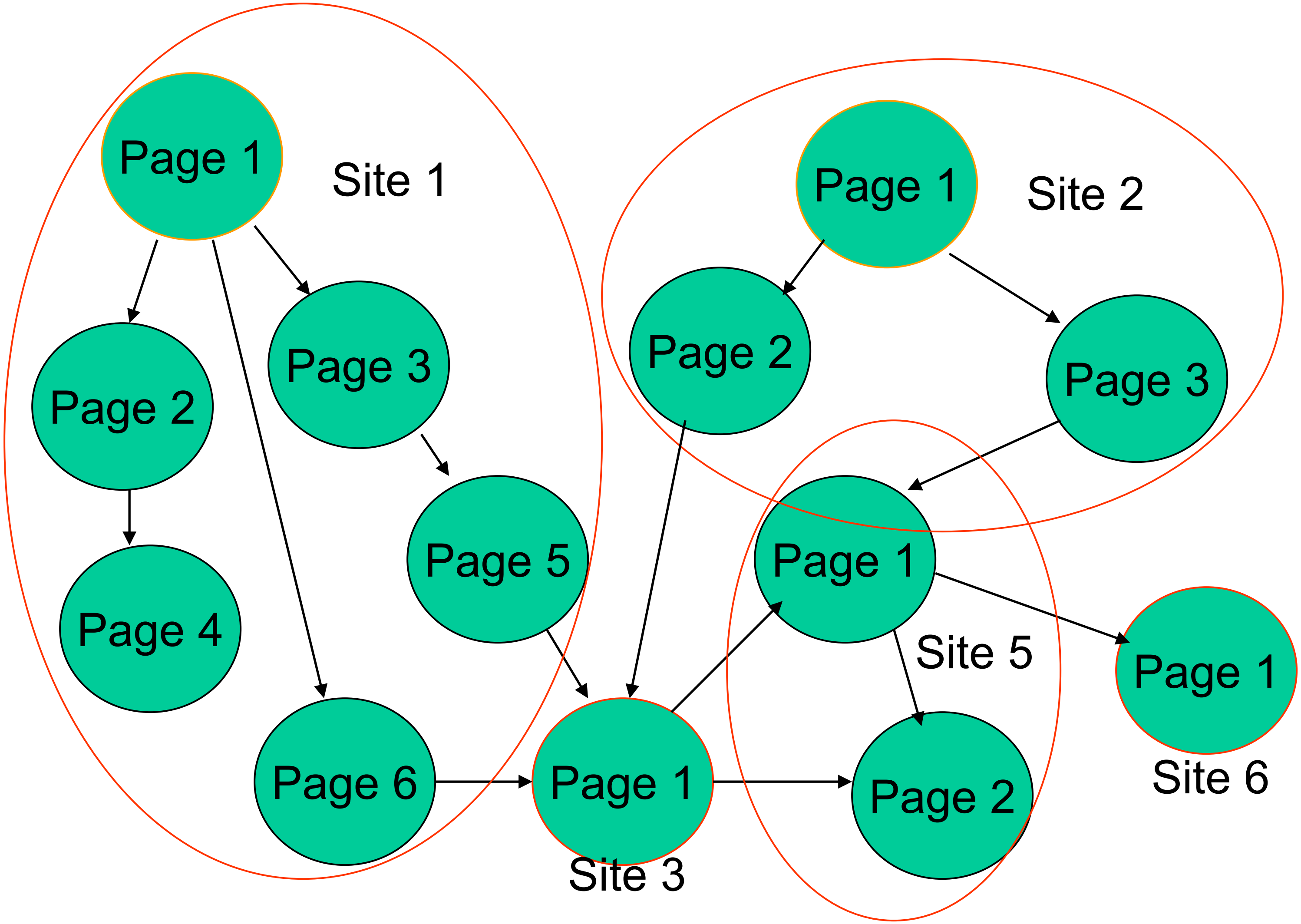
เอา query มาmatch กับ index

ทำ index ไว้

ค:40

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

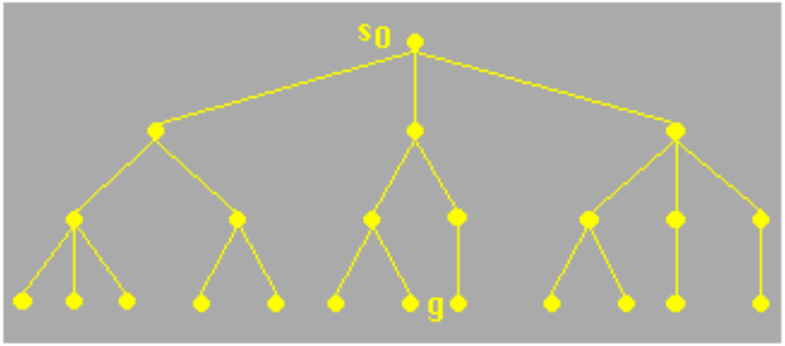
เยี่ยมweb แบ
ลึก



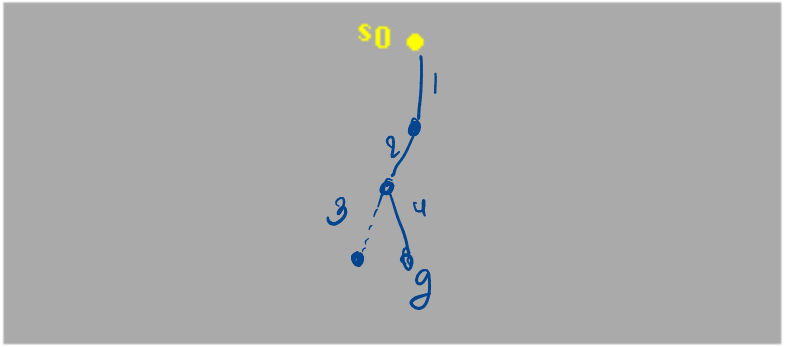
Site	Page
1	1
1	2
1	4
1	6
1	3
1	5
3	1
5	1
6	1
5	2
2	1
2	2
2	3

20:15

Depth-first search

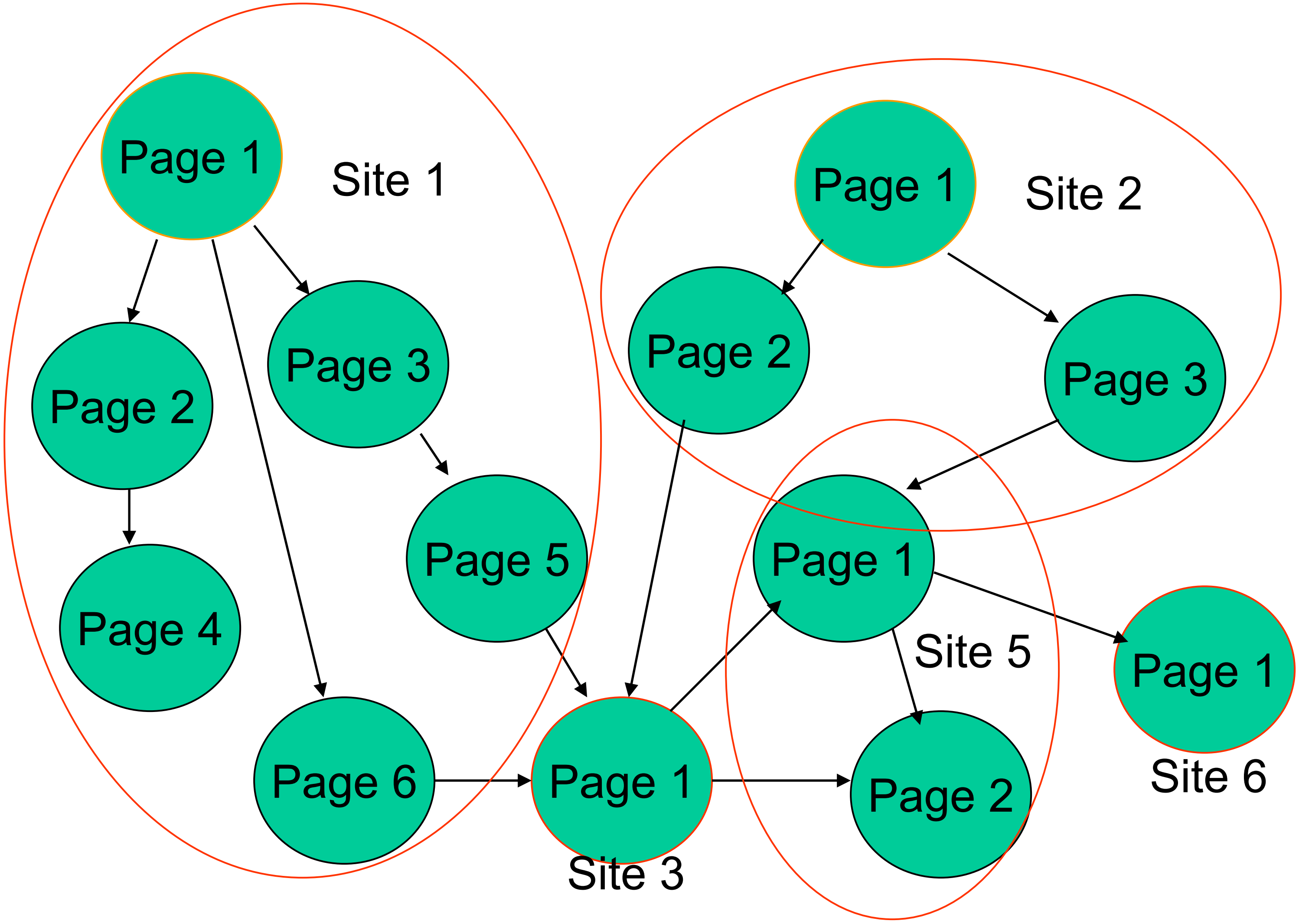


ต้องปรับ tree ให้ ผงาพ



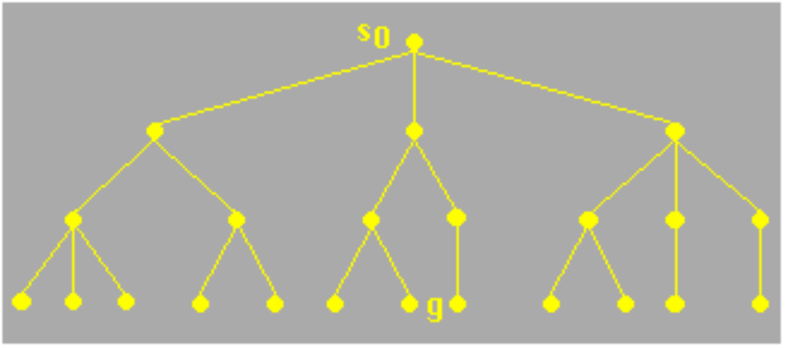
21:30

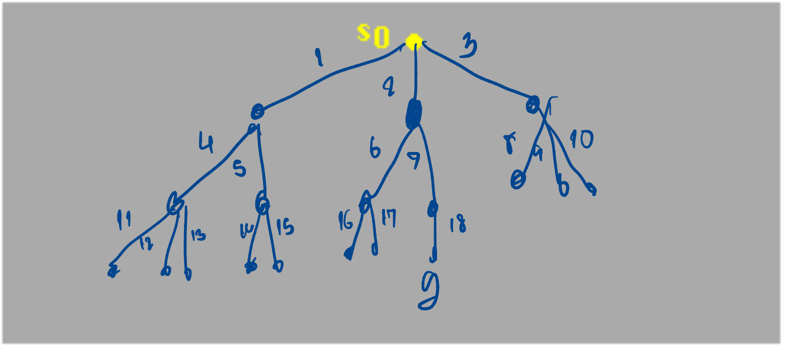
Breadth First Crawling (more complex – graphs & sites)



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





24:28

Centralized Architecture (Crawler-indexer)

ดี : ง่าย

Problem

1. Volumn of the data

Traffic ∝ data 2. Traffic (Crawler retrieve entire object) ทั้งก้อน- เท่า

3. High load at Web Servers

↓

มีแหล่งดึง ไม่การแบ่ง คือ Die

24:48

Distributed Architecture (Harvest)

Definition

1 Broker : หลาย/1 Gather

data ที่ม

แหล่งความรู้

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

→ มี index ตลอดเวลาว่า broker ไหนมีความรู้อะไร

query ตรง = ตอบเลย

ไม่ตรง = โยนไปหาคนอื่นที่ตรง (Broker อื่น)

2. Brokers หาย

หายหน้าแต่ละตัวมีแน่ๆ ความรู้ เป็นของตัวเอง มีหลาย

ของตัวเอง

จ.มี

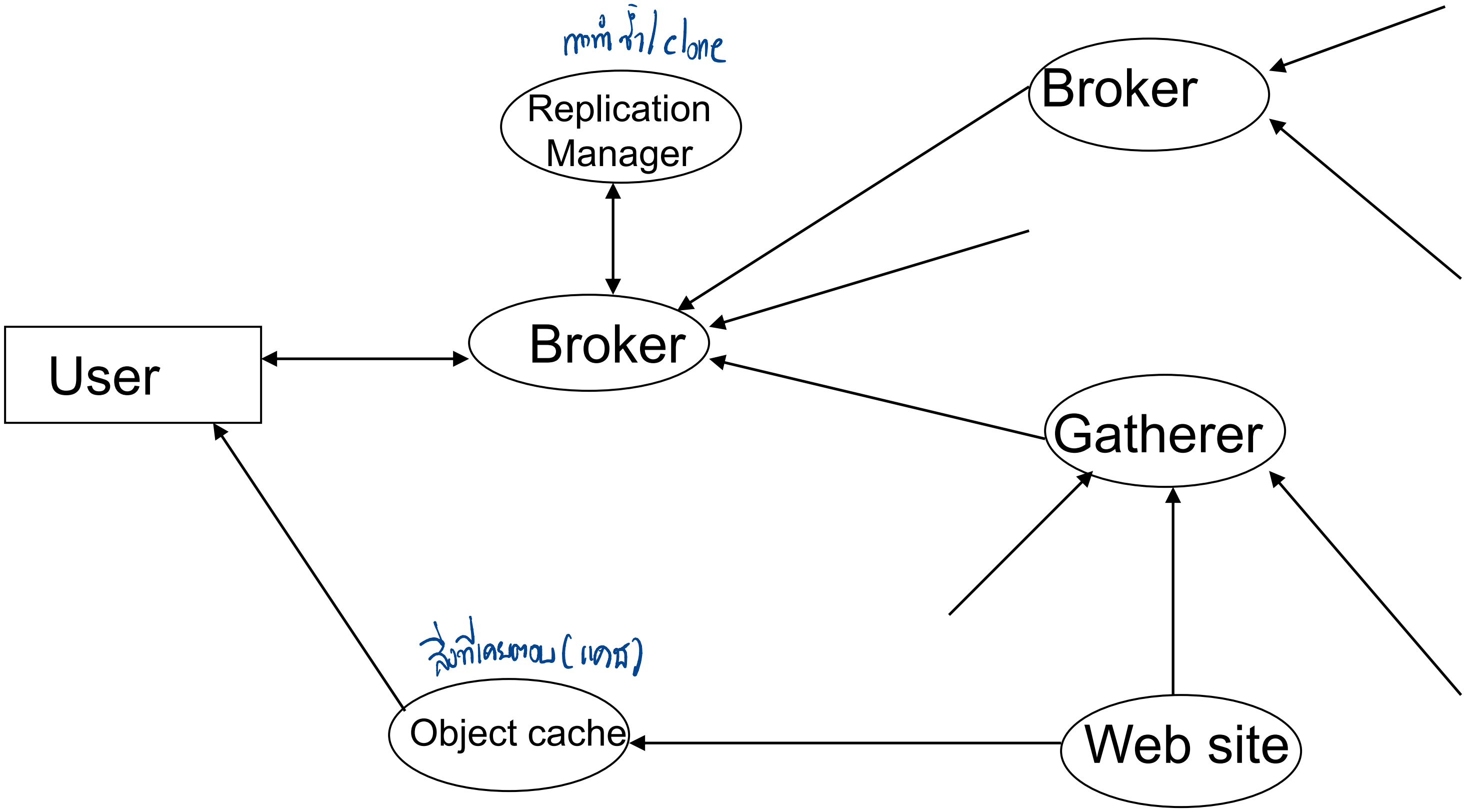
-Provide indexing mechanism and query interface to data gathered

index

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

29:01

Distributed Architecture (Harvest architecture)



การทำซ้ำ/clone

สิ่งที่เคยตอบ (แคช)