

Web Crawlers & Hyperlink Analysis

STBI 2010

Husni

husni.trunojoyo.ac.id

Ikhtisar

- Web Crawler
 - Sejarah & Definisi
 - Arsitektur
- Analisis Hyperlink
 - HITS
 - PageRank

Definisi

- Spiders, robots, bots, aggregators, agents and intelligent agents.
- An internet-aware program that can retrieve information from specific location on the internet
- A program that collects documents by recursively fetching links from a set of starting pages.
- Web crawlers are programs that exploit the graph structures of the web to move from page to page

Penelitian Crawler

- There is a few research about crawlers
- “...very little research has been done on crawlers.”

Junghoo Cho, Hector Garcia-Molina, Lawrence Page, *Efficient Crawling Through URL Ordering*, Stanford University , 1998

Penelitian Crawler

- “Unfortunately, many of the techniques used by dot-coms, and especially the resulting performance, are private, behind company walls, or are disclosed in patents....”

Arvind Arasu, et al, *Searching the web*, Stanford university 2001

- “... due to the competitive nature of the search engine business, the designs of these crawlers have not been publicly described. There are two notable exceptions : The Google crawler and the Internet Archive crawler. Unfortunately , the descriptions of these crawlers in the literature are too terse to enable reproducibility”

Alan Heydon and Marc Najork, *Mercator : A scalable, Extensible Web Crawler* , Compaq System Research Center 2001

Penelitian Crawler

- Web crawling and indexing companies are rather protective about the engineering details of their software assets. Much of the discussion of the typical anatomy of large crawlers is guided by an early paper discussing the crawling system for [26] Google , as well as a paper about the design of Mercator, a crawler written in Java at Compaq Research Center [108].

*Soumen Chakrabarti. **Mining The Web discovering knowledge from hypertext data**. Morgan Kaufmann 2003*

Penelitian Crawler

- 1993 : First crawler, Matthew Gray's Wanderer
- 1994 :
 - David Eichmann. The RBSE Spider – Balancing Effective Search Against Web Load. In *Proceedings of the First International World Wide Web Conference*, 1994.
 - Oliver A. McBryan. GENVL and WWW : Tools for taming the web. In *Proceedings of the First International World Wide Web Conference*, 1994.
 - Brian Pinkerton . Finding What people Want : Experiences with the webCrawler. In *Proceedings of the Second International World Wide Web Conference*, 1994.

Penelitian Crawler

- **1997 : www.archive.org crawler**

M. Burner. Crawling towards eternity: Building an archive of the world wide web. *Web Techniques Magazine*, 2(5), May 1997.

- **1998 : Google crawler**

S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of the 7th World Wide Web Conference*, pages 107–117, 1998.

- **1999 : Mercator**

A. Heydon and M. Najork. Mercator: A scalable, extensible web crawler. *World Wide Web*, 2(4):219–229, 1999.

Penelitian Crawler

- 2001 : WebFountain Crawler

J. Edwards, K. S. McCurley, and J. A. Tomlin. An adaptive model for optimizing performance of an incremental web crawler. In *Proceedings of the 10th International World Wide Web Conference*, pages 106-113, May 2001.

- 2002 :

- Cho and Garcia-Molina's crawler

J. Cho and H. Garcia-Molina. Parallel crawlers. In *Proceedings of the 11th International World Wide Web Conference*, 2002.

- UbiCrawler

P. Boldi, B. Codenotti, M. Santini, and S. Vigna. Ubicrawler: A scalable fully distributed web crawler. In *Proceedings of the 8th Australian World Wide Web Conference*, July 2002.

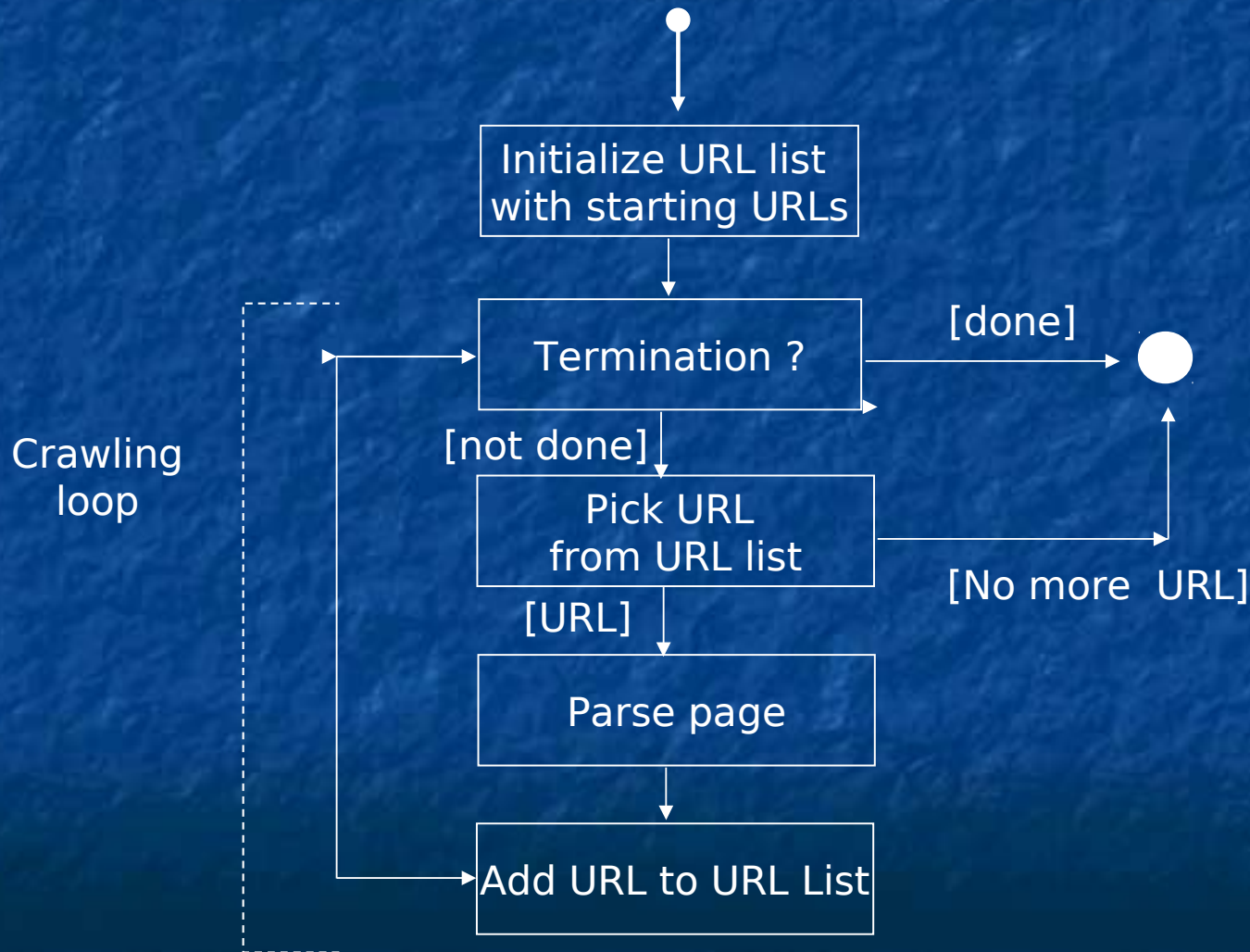
Penelitian Crawler

- 2002 : Shkapenyuk and Suel's Crawler
V. Shkapenyuk and T. Suel. Design and implementation of a high-performance distributed web crawler. In *IEEE International Conference on Data Engineering (ICDE)*, Feb. 2002.
- 2004 : Carlos Castillo
Castillo, C. Effective Web Crawling. Phd Thesis. University of Chile. November 2004.
- 2005 : DynaBot
Daniel Rocco, James Caverlee, Ling Liu, Terence Critchlow. Posters: Exploiting the Deep Web with DynaBot : Matching, Probing, and Ranking. *Special interest tracks and posters of the 14th international conference on World Wide Web*, May 2005
- 2006 : ?

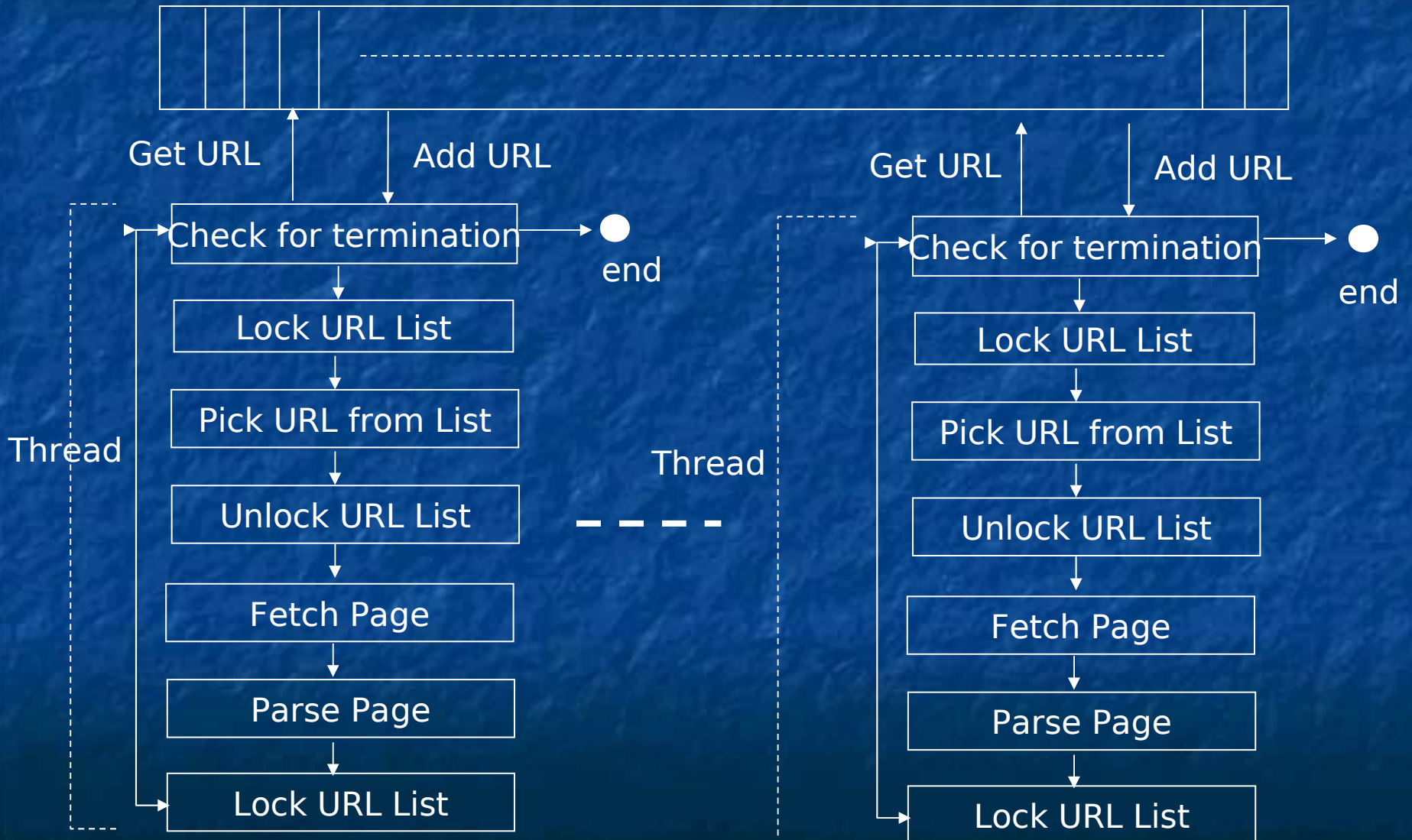
Algoritma Dasar Crawler

1. Remove a URL from the unvisited URL list
2. Determine the IP Address of its host name
3. Download the corresponding document
4. Extract any links contained in it.
5. If the URL is new, add it to the list of unvisited URLs
6. Process the downloaded document
7. Back to step 1

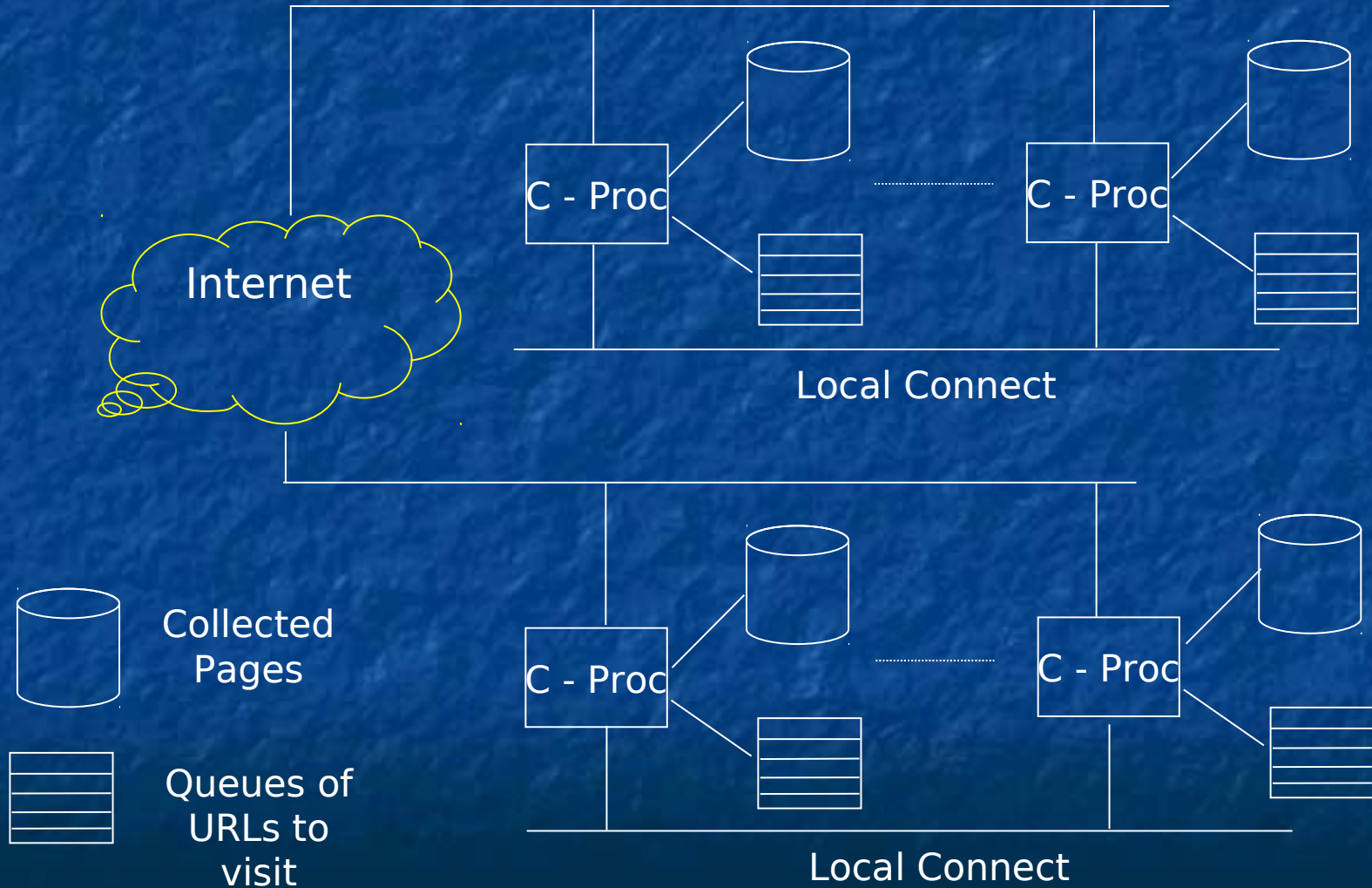
Crawler Tunggal



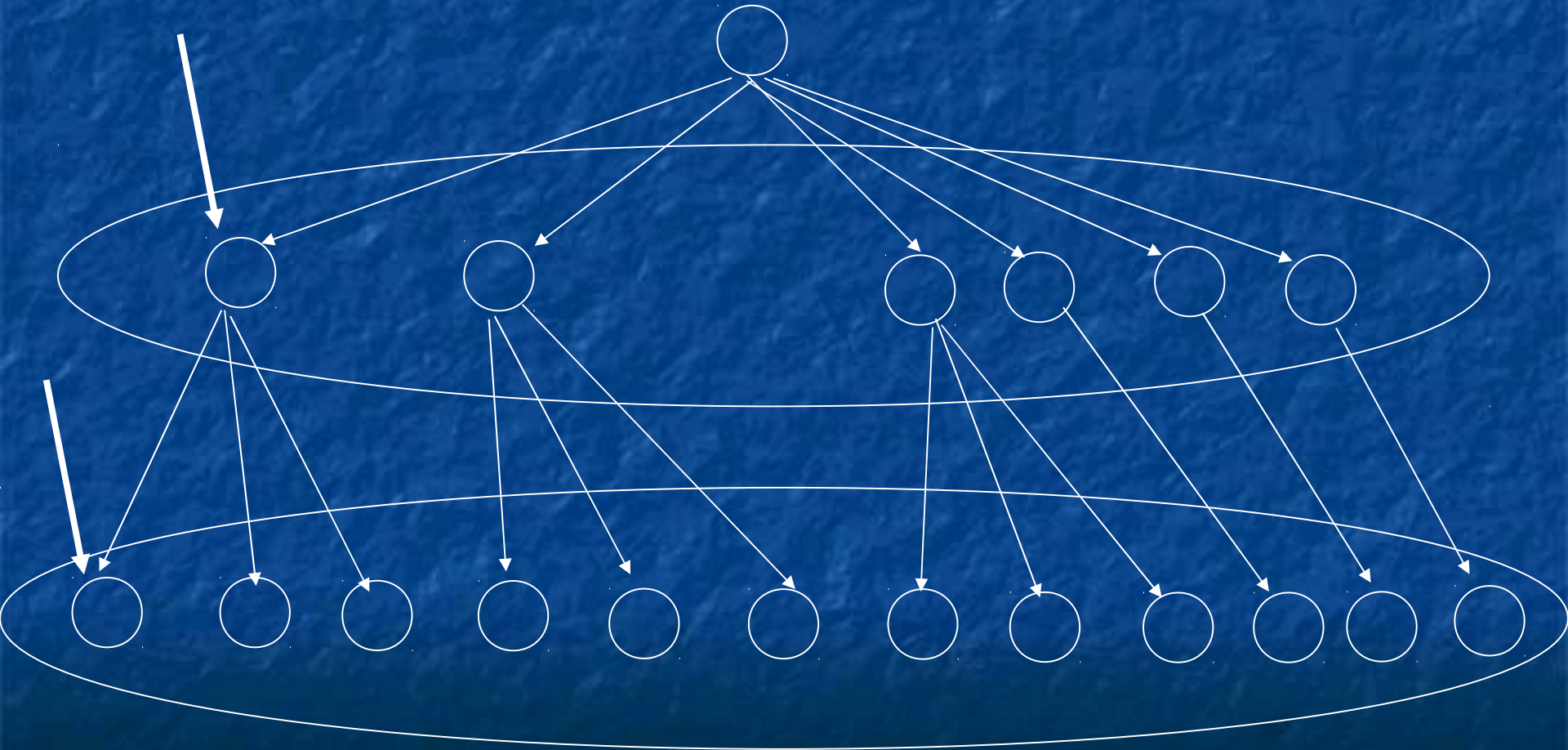
Crawler Banyak Thread



Crawler Paralel



Breadth-First Search



Protocol Robot

- Contains part of the web site that a crawler should not visit.
- Placed at the root of a web site, robots.txt

```
# robots.txt for http://somehost.com/
```

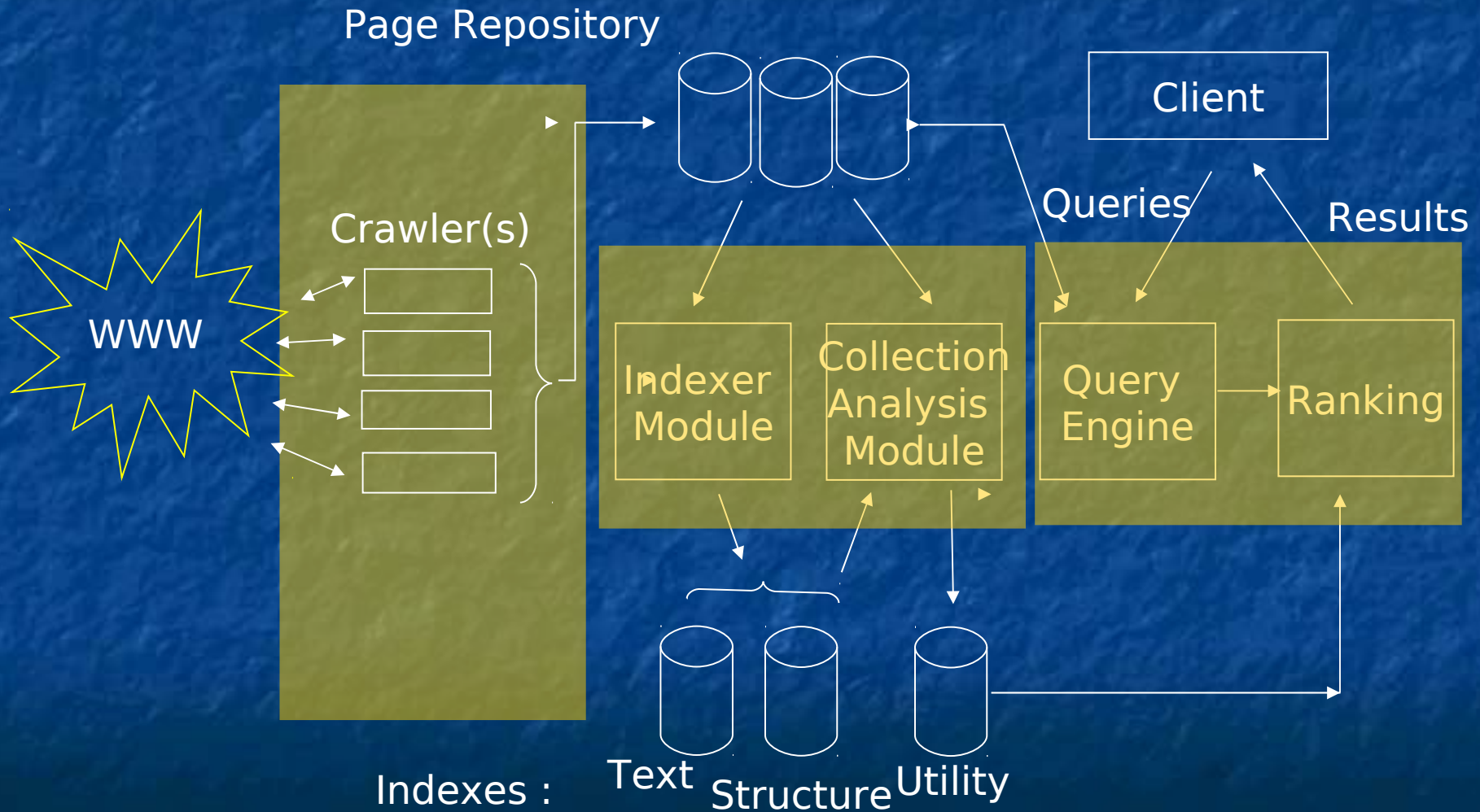
```
User-agent: *
```

```
Disallow: /cgi-bin/
```

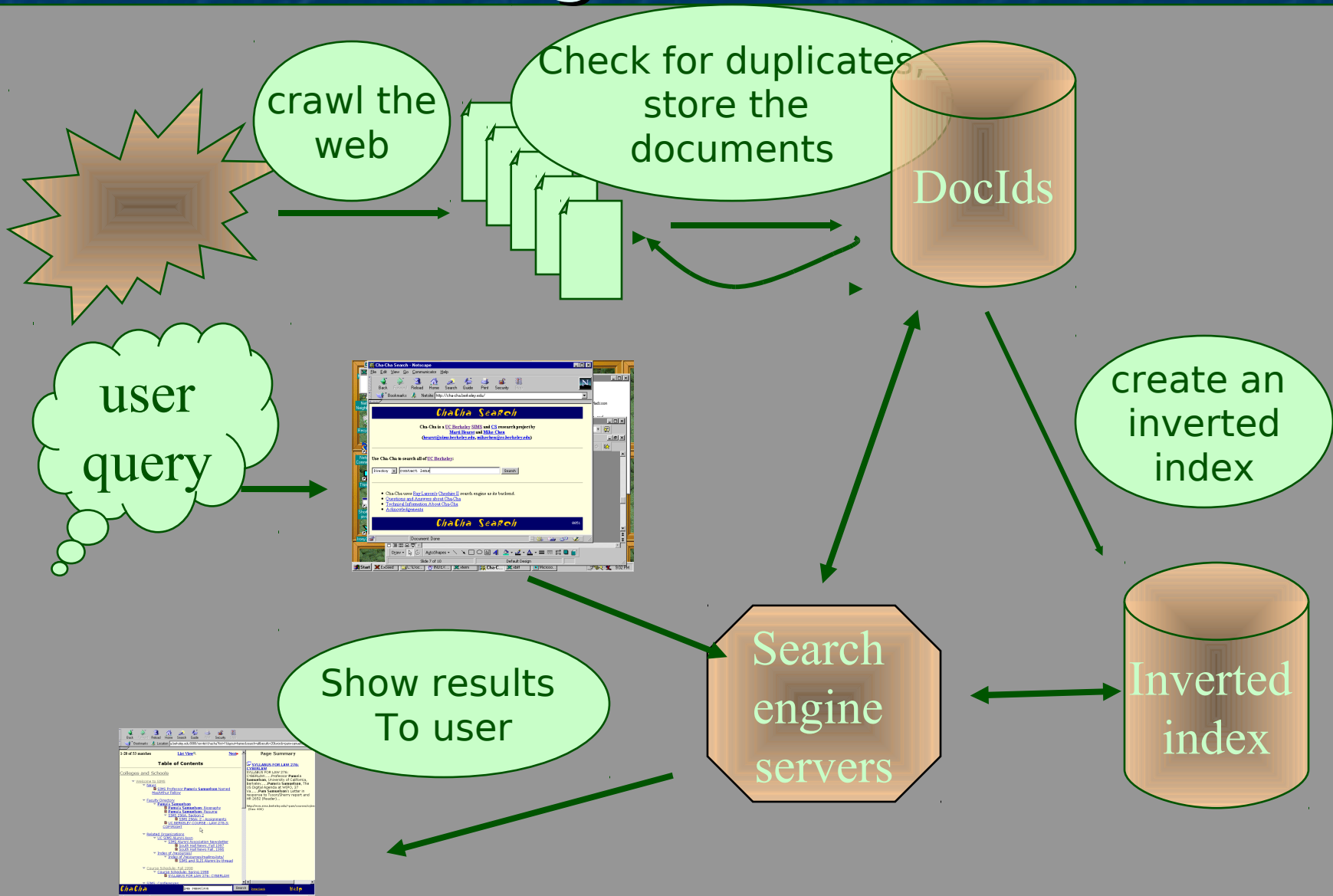
```
Disallow: /registration # Disallow robots on registration page
```

```
Disallow: /login
```

Search Engine: Architecture



Search Engine: Architecture



Search Engine: Komponen Utama

■ Crawlers

Collects documents by recursively fetching links from a set of starting pages.

Each crawler has different policies

The pages indexed by various search engines are different

■ The Indexer

Processes pages, decide which of them to index, build various data structures representing the pages (inverted index, web graph, etc), different representation among search engines.

Might also build additional structure (LSI)

■ The Query Processor

Processes user queries and returns matching answers in an order determined by a ranking algorithm.

Isu Seputar Crawler

1. General architecture
2. What pages should the crawler download ?
3. How should the crawler refresh pages ?
4. How should the load on the visited web sites be minimized ?
5. How should the crawling process be parallelized ?

Analisis Halaman Web

- Content-based analysis
 - Based on the words in documents
 - Each document or query is represented as a term vector
 - E.g : Vector space model algorithm, tf-idf
- Connectivity-based analysis
 - Use hyperlink structure
 - Used to indicate “importance” measure of web pages
 - E.g : PageRank, HITS

Analisis Hyperlink

- Exploiting hyperlink structure of web pages to find relevant and importance pages for a user query
- Assumptions :
 1. Hyperlink from page A to page B is a recommendation of page B of the author of page A
 2. If page A and page B are connected by a hyperlink , then might be on the same topic.
- Used for crawling, ranking, computing the geographic scope of a web page, finding mirrored hosts , computing statistics of web pages and search engines, web page categorization.

Analisis Hyperlink

- Most popular methods :
 - HITS (1998)
(Hypertext Induced Topic Search)
By Jon Kleinberg
 - PageRank (1998)
By Lawrence Page & Sergey Brin
Google's founders

HITS

- Mencakup dua langkah:
 1. Building a neighborhood graph N related to the query terms
 2. Computing authority and hub scores for each document in N , and present the two ranked list of the most authoritative and most “hubby” documents

HITS

freedom : term 1 - doc 3, doc 117, doc 3999

.

.

registration : term 10 - doc 15, doc 3, doc 101,
doc 19,

doc 1199, doc 280

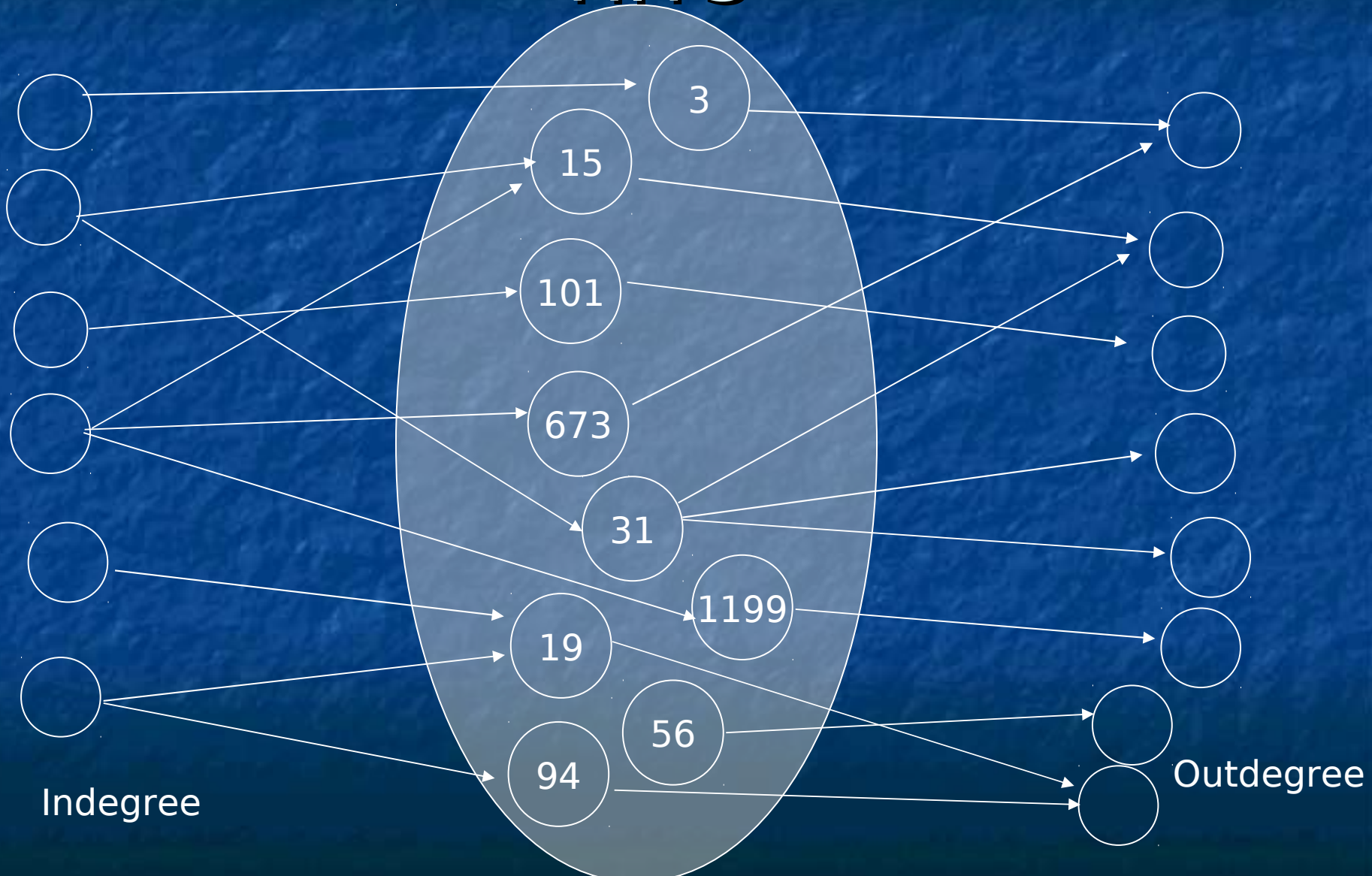
faculty : term 11 - doc 56, doc 94, doc 31, doc 3

.

.

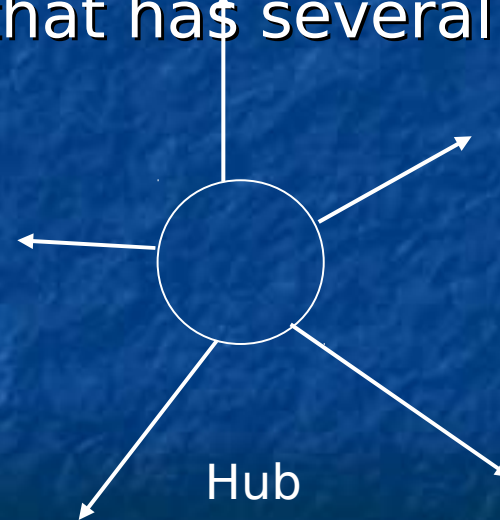
graduation : term m - doc 223

HITS



HITS

- HITS defines Authorities and Hubs
- An authority is a document with several inlinks
- A hub is a document that has several outlinks



Perhitungan HITS

- Good authorities are pointed to by good hubs
- Good hubs point to good authorities
- Page i has both authority score x_i and a hub score y_i

- $x_i^{(k)} = \sum_{e_{ij} \in E} y_j^{(k-1)}$ For $k = 1, 2, 3, \dots$
- $y_i^{(k)} = \sum_{e_{ij} \in E} x_j^{(k)}$
- E = the set of all directed edges in the web graph
- e_{ij} = the directed edge from node i to node j
- Given initial authority score $x_i^{(0)}$ and hub score $y_i^{(0)}$

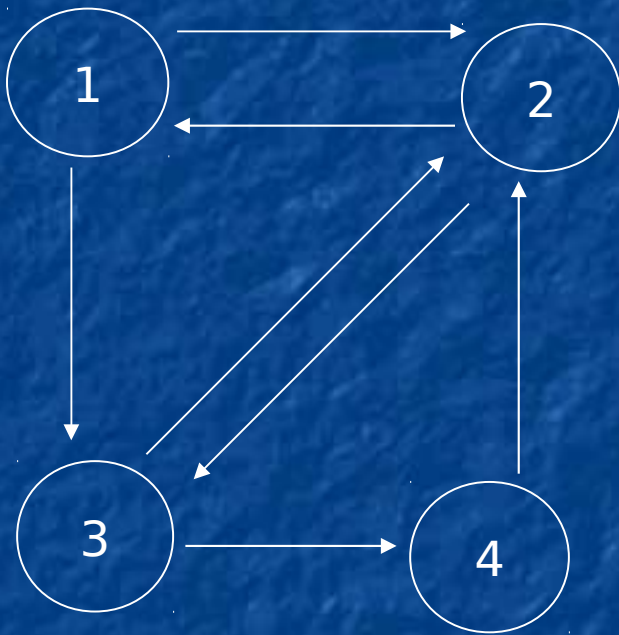
Perhitungan HITS

$$X_i^{(k)} = \sum_{j: e_{ji} \in E} Y_j^{(k-1)} \quad Y_i^{(k)} = \sum_{j: e_{ij} \in E} X_j^{(k)} \quad \text{For } k = 1, 2, 3, \dots$$

- Can be written in matrix L of the directed web graph

$L_{ij} = 1$, there exists an edge from node i to node j
 0 , otherwise

Perhitungan HITS



$$L = \begin{matrix} & d1 & d2 & d3 & d4 \\ d1 & \begin{pmatrix} 0 & 1 & 1 & 0 \end{pmatrix} \\ d2 & \begin{pmatrix} 1 & 0 & 1 & 0 \end{pmatrix} \\ d3 & \begin{pmatrix} 0 & 1 & 0 & 1 \end{pmatrix} \\ d4 & \begin{pmatrix} 0 & 1 & 0 & 0 \end{pmatrix} \end{matrix}$$

$$X_i^{(k)} = L^T Y_j^{(k-1)}$$

$$\text{And } Y_i^{(k)} = L X_j^{(k)}$$

Perhitungan HITS

1. Initialize $y^{(0)} = e$, e is a column vector of all ones

2. Until convergence do

$$x^{(k)} = L^T y^{(k-1)}$$

$$y^{(k)} = L x^{(k)}$$

$$k = k + 1$$

$$x^{(k)} = L^T L x^{(k-1)}$$

$$y^{(k)} = L L^T y^{(k-1)}$$

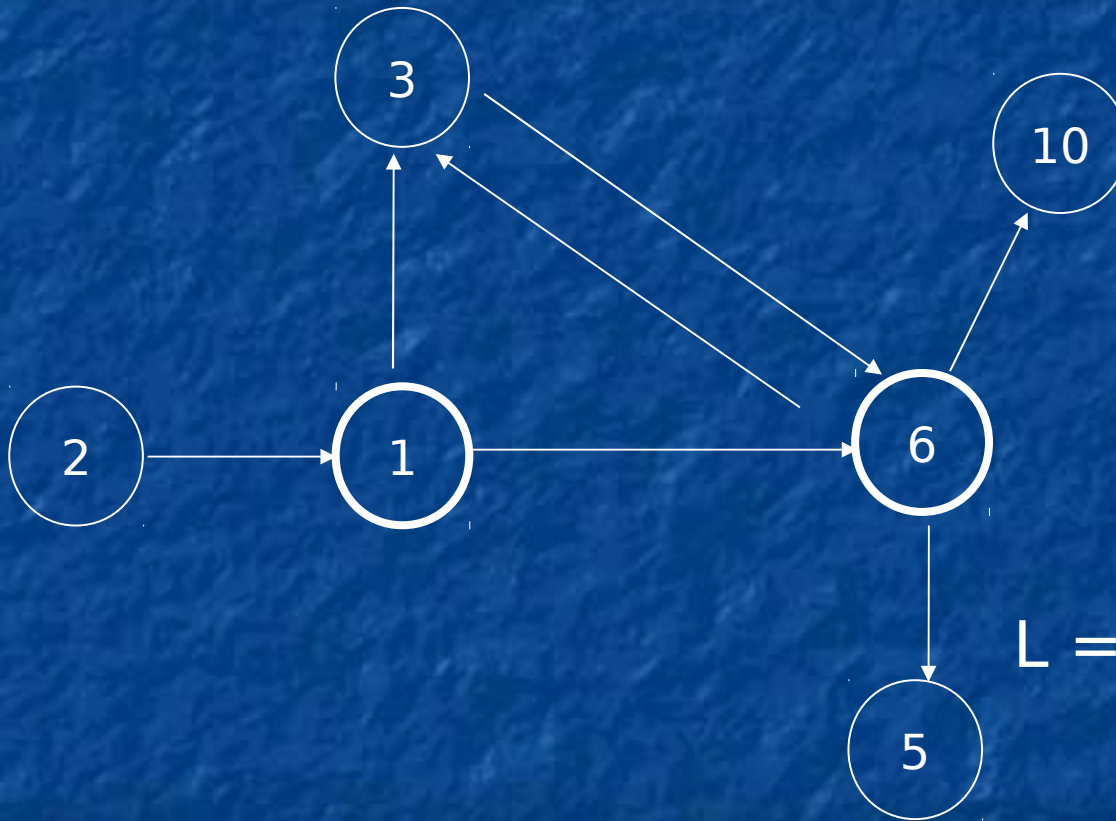
$L^T \mathbf{1}$ = authority matrix

Normalize $x^{(k)}$ and $y^{(k)}$

$L L^T$ = hub matrix

Computing authority vector X and hub vector Y can be viewed as finding dominant right-hand eigenvectors of $L^T L$ and $L L^T$

Contoh HITS



$L =$

	1	2	3	5	6	10
1	0	0	1	1	0	0
2	1	0	0	0	0	0
3	0	0	0	0	1	0
5	0	0	0	0	0	0
6	0	0	1	1	0	0
10	0	0	0	0	1	0

Contoh HITS

$$L^T L = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 5 & 6 & 10 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 5 \\ 6 \\ 10 \end{matrix} & \begin{pmatrix} 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{matrix}$$

$$L L^T = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 5 & 6 & 10 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 5 \\ 6 \\ 10 \end{matrix} & \begin{pmatrix} 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{matrix}$$

Authorities and Hub matrices

Contoh HITS

The normalized principles eigenvectors with the Authority score x and Hub y are :

$$X^T = (0 \quad 0 \quad .3660 \quad .1340 \quad .5 \quad 0)$$

$$Y^T = (.3660 \quad 0 \quad .2113 \quad 0 \quad .2113 \quad .2113)$$

$$\text{Authority Ranking} = (6 \quad 3 \quad 5 \quad 1 \quad 2 \quad 10)$$

$$\text{Hub Ranking} = (1 \quad 3 \quad 6 \quad 10 \quad 2 \quad 5)$$

Kekuatan & Kelemahan HITS

- Strengths
 - Dual rankings
- Weaknesses
 - Query-dependence
 - Hub score can be easily manipulated
 - It is possible that a very authoritative yet off-topic document be linked to a document containing the query terms (Topic drift)

PageRank

- Is a numerical value that represent how important a page is.
- casts a “vote” to page that it links to, the more vote cast to a page the more important the page.
- The importance of a page that links to a page determines how importance the link is.
- The importance score of a page is calculated from the vote cast for that page.
- Used by Google

PageRank

$$PR(A) = (1 - d) \times d [PR(t_1)/C(t_1) + PR(t_2)/C(t_2) + \dots + PR(t_n)/C(t_n)]$$

Where :

PR(A) = PageRank of page A

d = damping factor , usually set to 0.85

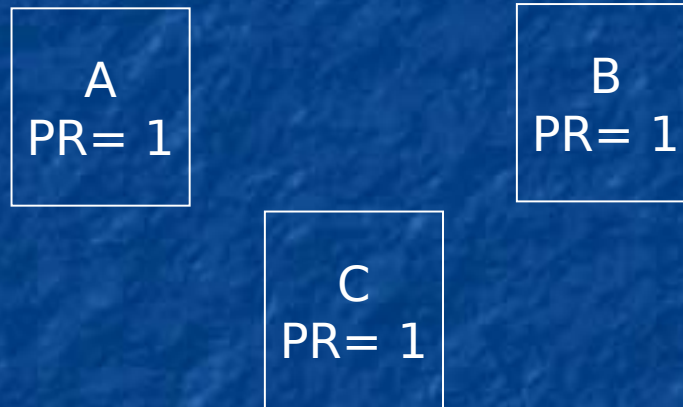
t1, t2, t3, ... tn = pages that link to page A

C() = the number of outlinks of a page

In a simpler way :

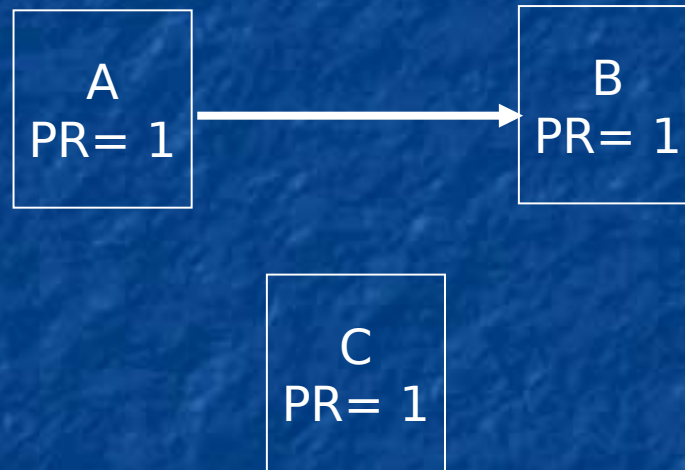
$$PR(A) = 0.15 \times 0.85 [\text{a share of the PageRank of every page that links to}]$$

Contoh PageRank



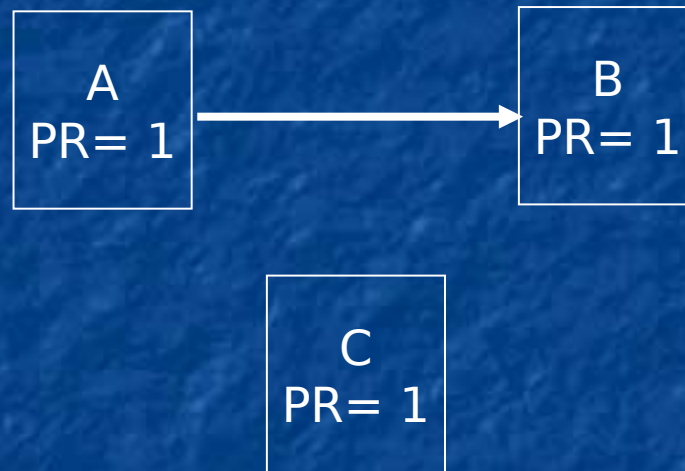
- Each page is assigned an initial PageRank of 1
- The site's maximum PageRank is 3
- $PR(A) = 0.15$
- $PR(B) = 0.15$
- $PR(C) = 0.15$
- The total PageRank in the site = 0.45, seriously wasting most of its potential PageRank

Contoh PageRank



- $PR(A) = 0.15$
- $PR(B) = 1$
- $PR(C) = 0.15$
- Page B's PageRank increase, because page A has "voted" for page B

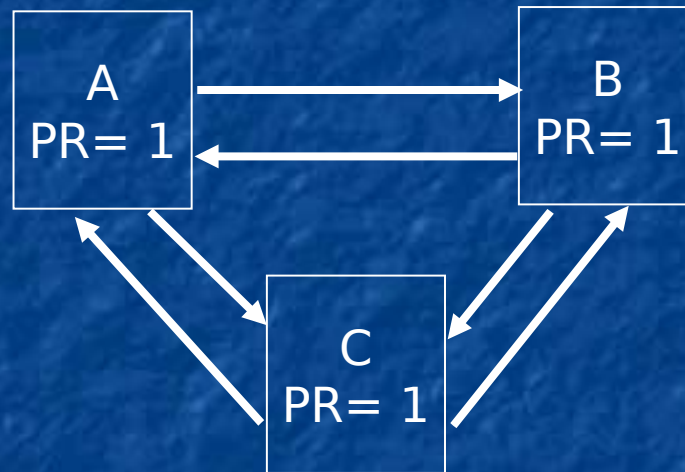
Contoh PageRank



After 100 iteration

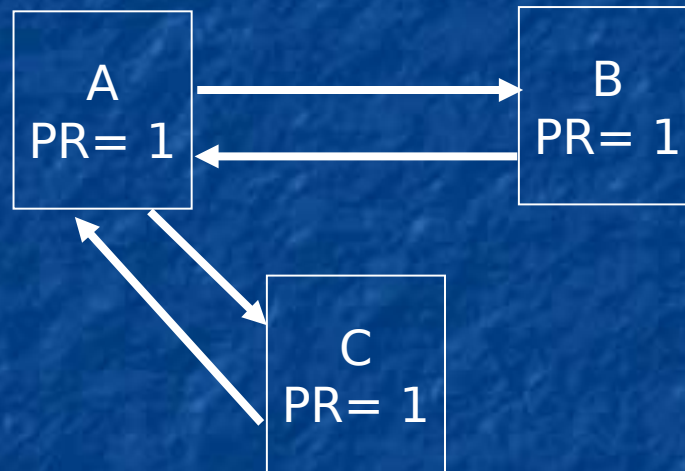
- $PR(A) = 0.15$
- $PR(B) = 0.2775$
- $PR(C) = 0.15$
- The total PageRank in the site 0.5775

Contoh PageRank



- No matter how many iterations are run, each page always end up with $PR = 1$
- $PR(A) = 1$
- $PR(B) = 1$
- $PR(C) = 1$
- This occur by linking in a loop

Contoh PageRank



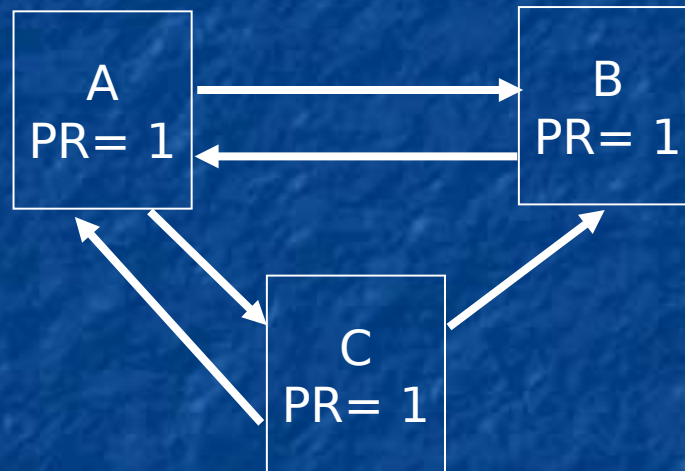
- $PR(A) = 1.85$
- $PR(B) = 0.575$
- $PR(C) = 0.575$

After 100 iterations :

- $PR(A) = 1.459459$
- $PR(B) = 0.7702703$
- $PR(C) = 0.7702703$

- The total pageRank is 3 (max), so none is being wasted
- Page A has a higher PR

PageRank Example



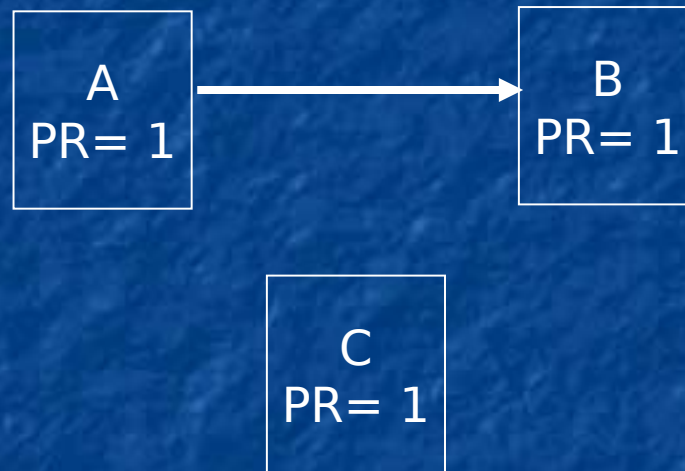
- $PR(A) = 1.425$
- $PR(B) = 1$
- $PR(C) = 0.575$

After 100 iterations :

- $PR(A) = 1.298245$
- $PR(B) = 0.999999$
- $PR(C) = 0.7017543$

- Page C share its “vote” between A and B
- Page A lost some values

Dangling Links



- Is a link to a page that has no links going from it or links to a page that has not been indexed
- Google removes the link shortly after the calculations start and reinstates them shortly before the calculations are finished.

PageRank Demo

- <http://homer.informatics.indiana.edu/cgi-bin/pagerank/cleanup.cgi>

PageRank Implementation

freedom : term 1 - doc 3, doc 117, doc 3999

.

.

registration : term 10 - doc 101, doc 87, doc 1199

faculty : term 11 - doc 280, doc 85

.

.

graduation : term m - doc 223

PageRank Implementation

- Query result on term 10 and 11 is
 $\{101, 280, 85, 87, 1199\}$
- $PR(87) = 0.3751$
 $PR(85) = 0.2862$
 $PR(101) = 0.04151$
 $PR(280) = 0.03721$
 $PR(1199) = 0.0023$
- Document 87 is the most important of the relevant documents

Strength and weaknesses of PageRank

- Weaknesses
 - The topic drift problem due to the importance of determining an accurate relevancy score.
 - Much work, thought and heuristic must be applied by Google engineers to determine the relevancy score, otherwise, the PageRank retrieved list might often be useless to a user.
 - Question : why does importance serve as such a good proxy to relevance ?
 - Some of these questions might be unanswered due to the proprietary nature

Strength and weaknesses of PageRank

- Strengths
 - The use of importance rather than relevance
 - By measuring importance, query-dependence is not an issue.
 - Query-independence
 - Faster retrieval

HITS vs PageRank

HITS	PageRank
Connectivity-based analysis	Connectivity-based analysis
Eigenvector & eigenvalues calculation	Eigenvector & eigenvalues calculation
Query-dependence	Query-independence
Relevance	Importance

Q & A

Thank you