SNS COLLEGE OF ENGINEERING

Kurumbapalayam (Po), Coimbatore – 641 107

An Autonomous Institution

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A' Grade

Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

COURSE NAME : 19CS732 INFORMATION RETRIEVAL TECHNIQUES

IV YEAR / VII SEMESTER

Unit 4- WEB RETRIEVAL AND WEB CRAWLING

Topic 1 : The Web Search Engine Architectures and Cluster based Architecture

SVM Classifier - **Problem**

- Issues
- The web is really infinite
- Dynamic content, e.g., calendars
- Soft 404: www.yahoo.com/<anything> is a valid page
- Static web contains syntactic duplication, mostly due to mirroring (~30%)
- Some servers are seldom connected
- Who cares?
- Media, and consequently the user
- Engine design
- Engine crawl policy. Impact on recall.

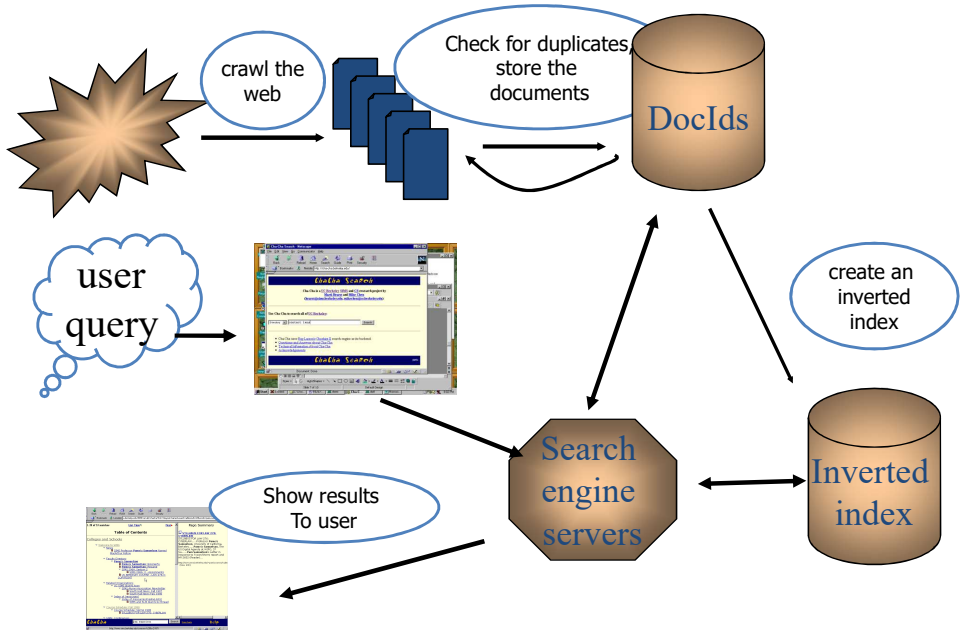
The Web Search Engine Architectures

- Unedited – anyone can enter
  - Quality issues
  - Spam
- Varied information types
  - Phone book, brochures, catalogs, dissertations, news reports, weather, all in one place!
- Different kinds of users
  - Online catalogs
    - scholars searching scholarly literature
  - Web
    - Every type of person with every type of goal
- Scale
  - Hundreds of millions of searches/day; billions of docs

Directories Vs Search Engines

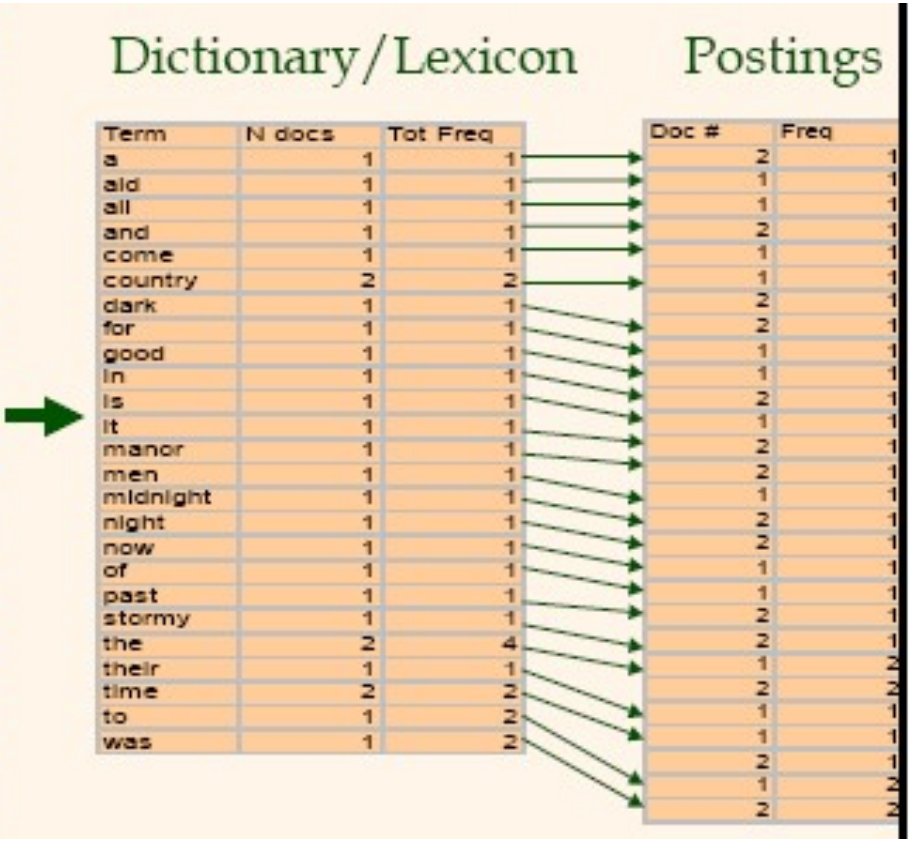
- Directories
  - Hand-selected sites
  - Search over the contents of the descriptions of the pages
  - Organized in advance into categories
- Search Engines
  - All pages in all sites
  - Search over the contents of the *pages themselves*
  - Organized after the query by relevance rankings or other scores

Standard Web Search Engine Architecture



How Inverted Files are Created?

Term	Doc #	Freq
a	2	1
aid	1	1
all	1	1
and	2	1
come	1	1
country	1	1
country	2	1
dark	2	1
for	1	1
good	1	1
in	2	1
is	1	1
it	2	1
manor	2	1
men	1	1
midnight	2	1
night	2	1
now	1	1
of	1	1
past	2	1
stormy	2	1
the	1	2
the	2	2
their	1	1
time	1	1
time	2	1
to	1	2
was	2	2



Inverted indexes

- Permit fast search for individual terms
- For each term, you get a list consisting of:
  - document ID
  - frequency of term in doc (optional)
  - position of term in doc (optional)
- These lists can be used to solve Boolean queries:
  - country -> d1, d2
  - manor -> d2
  - country AND manor -> d2
- Also used for statistical ranking algorithms

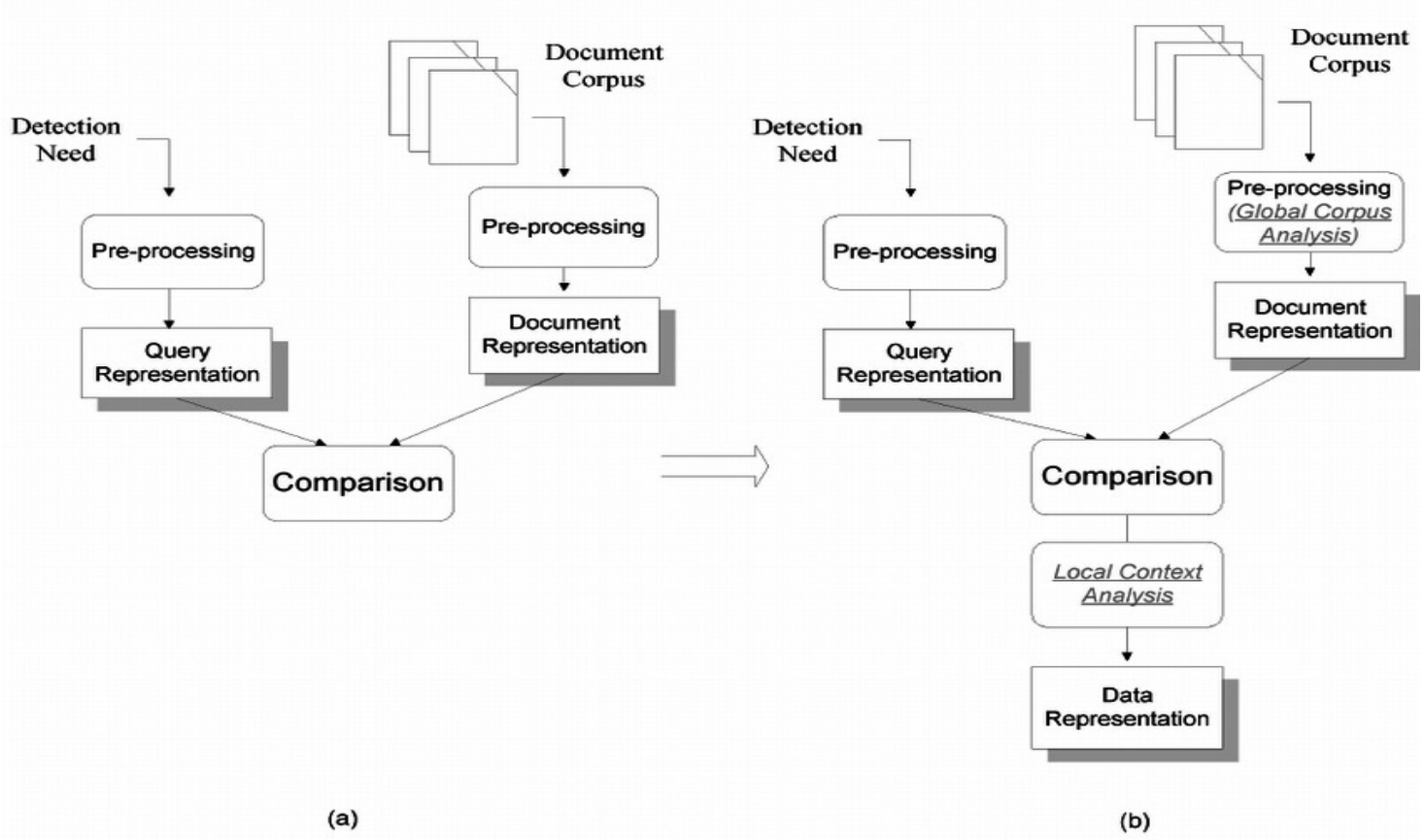
Cluster based Architecture

- Document clustering
  - Motivations
  - Document representations
  - Success criteria
- Clustering algorithms
  - Partitional
  - Hierarchical

What is clustering?

- Clustering: the process of grouping a set of objects into classes of similar objects
  - Documents within a cluster should be similar.
  - Documents from different clusters should be dissimilar.
- The commonest form of *unsupervised learning*
  - Unsupervised learning = learning from raw data, as opposed to supervised data where a classification of examples is given
  - A common and important task that finds many applications in IR and other places

Clustering Architecture

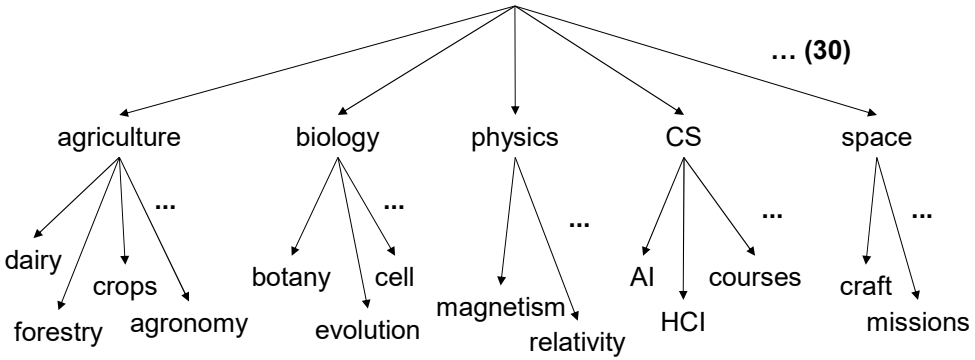


Clustering - Cont..

- **Whole corpus analysis/navigation**
  - **Better user interface: search without typing**
- For improving recall in search applications
  - Better search results (like pseudo RF)
- For better navigation of search results
  - Effective "user recall" will be higher
- For speeding up vector space retrieval
  - Cluster-based retrieval gives faster search

Yahoo! Hierarchy *isn't* clustering but *is* the kind of output you want from clustering

www.yahoo.com/Science



Activity

Disadvantages

- Cost is high. Since the cluster needs good hardware and a design, it will be costly comparing to a non-clustered server management design. Being not cost effective is a main disadvantage of this particular design.
- Since clustering needs more servers and hardware to establish one, monitoring and maintenance is hard. Thus increase the infrastructure.

Advantages

- Clustering servers is completely a scalable solution. You can add resources to the cluster afterwards.
- If a server in the cluster needs any maintenance, you can do it by stopping it while handing the load over to other servers.
- Among high availability options, clustering takes a special place since it is reliable and easy to configure. In case of a server is having a problem providing the services furthermore, other servers in the cluster can take the load.

Assessment 1

1. List out the Advantages of clustering Architecture
   a)_____
   b)_____
   c)_____
   d)_____

2. Identify the disadvantages of clustering Architecture
   a)_____
   b)_____
   c)_____
   d)_____

TEXT BOOKS:

1. Ricardo Baeza-Yates and Berthier Ribeiro-Neto, —Modern Information Retrieval: The Concepts and Technology behind Search, Second Edition, ACM Press Books, 2011.

2. Ricci, F, Rokach, L. Shapira, B.Kantor, —Recommender Systems Handbook||, First Edition, 2011.

REFERENCES:

1. C. Manning, P. Raghavan, and H. Schütze, —Introduction to Information Retrieval, Cambridge University Press, 2008.

2. Stefan Buettcher, Charles L. A. Clarke and Gordon V. Cormack, —Information Retrieval: Implementing and Evaluating Search Engines, The MIT Press, 2010.

THANK YOU