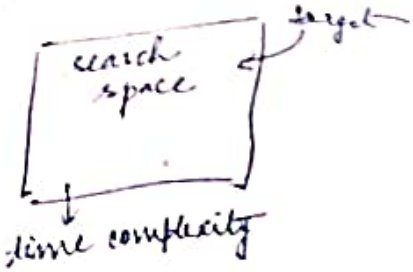
30.09.21

Search Technique.

What is searching? — target in a search space



Eg: - linear
- binary.

Storage

i) **Alphabetical order** (Dictionary arrangement)
- The data are arrange in alphabetical

ii) **Numerical Arrangement.**
- ascending / descending order

iii) **Classified arrangement**
- using categories

Steps that involve in search process.

i) Recognize & state the need.
ii) Development of search strategy ← efficient method (eg: ...)
iii) Execution of search strategy
iv) Review search results
v) Edit search results
vi) Evaluation & feedback.

Search string → structure
- i) Syntactic value
- ii) Semantic value
- iii) Boolean operator → OR, AND, NOT
  - meaning

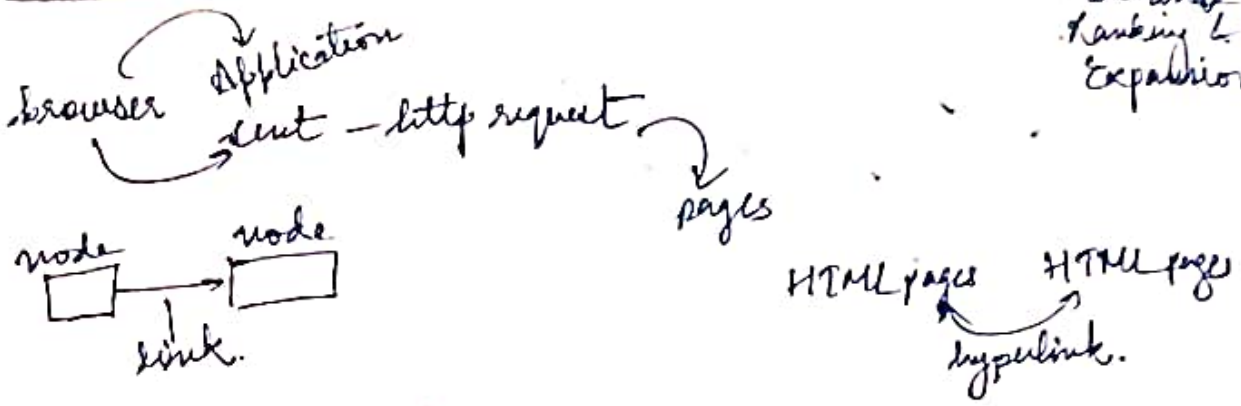
Different types of search

- i) Field based search → DB (eg: SQL) query
- ii) Full text search
- iii) Truncation search → search is conducted diff. form of word having the same root.
- iv) Proximity search
- v) Limiting search
- vi) Range search
- vii) Simple search.
- viii) Advance search

01/10/24.

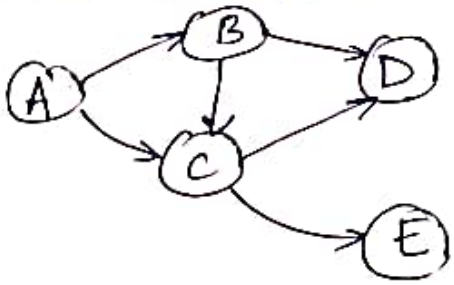
Unit 4 -
Advance topic
Searching
Relevance feedback
Ranking & Scoring
Expansion of

Web Search basics.



WEB GRAPH.

It consists of HTML pages together with the hyperlink represented as directed graph where each web page is a node & each hyperlink as a directed edge.



Indegree.
C = 2
B = 1
D = 2

POWER LAW

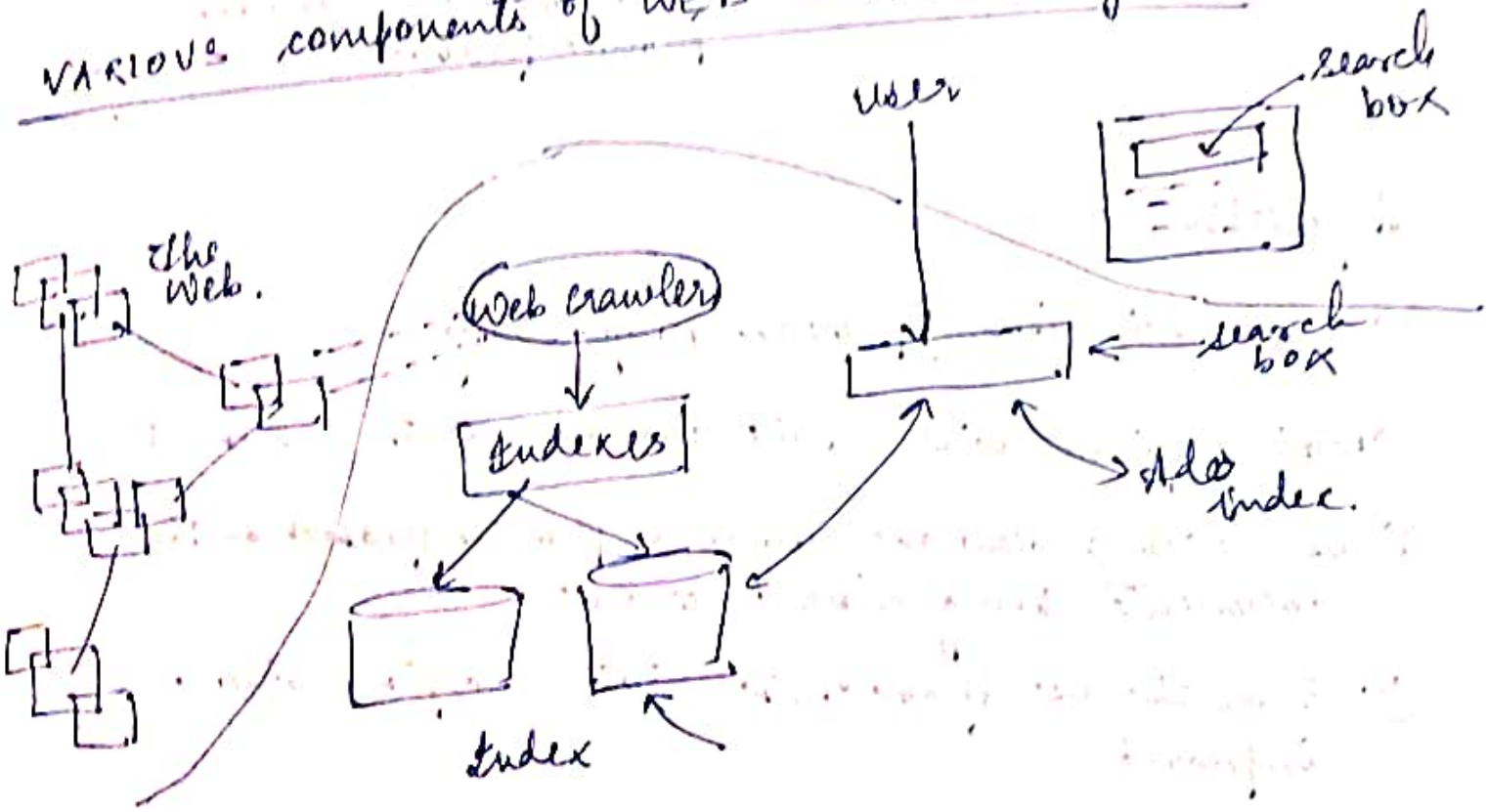
Total no. of web pages with indegree is proportional to $\frac{1}{i^{\alpha}}$, α is Ziff's law of distribution of word.

01.10.24

BOWTIE STRUCTURE OF WEB

Spam — advertisements
- unwanted e-mail / junk sent in bulk.

VARIOUS components of WEB SEARCH engine

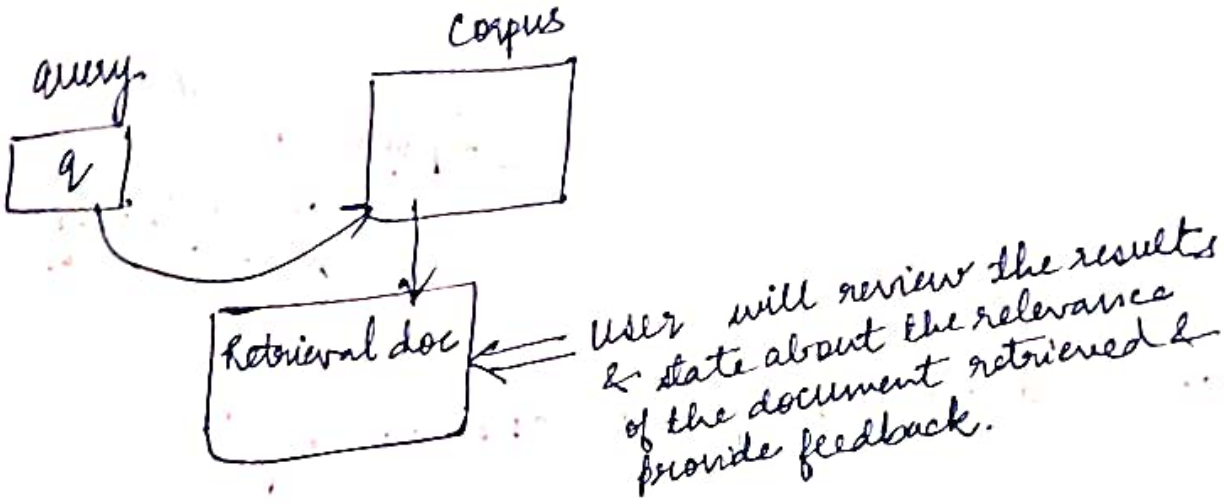


07/10/24

~~Relevance feed~~

RELEVANCE FEEDBACK

(1) Issue a query string



Algorithm:

(1) User will issue/provide query string

(2) Retrieval system will result the documents as relevant

(3) user will review the results & provide feedback on the documents being relevant or not.

(4) From the user feedback, the retrieval system will be improved.

07/10/24.

Near duplicates & Shingling. (Prev. chap. — topic was missed)

Multiple web pages → any 2 duplicates

same content ← diff. webpages

Shingling ~~technique~~

↓

mechanism / tech.

↓

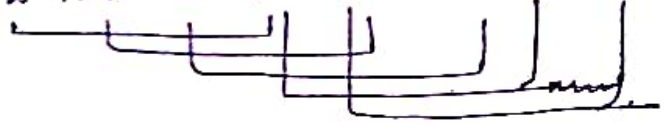
identify the near duplicates web pages.

Given: an integer k & sequence of terms in document d

shingling of document d ⇒ set of all consecutive sequence of k terms in document.

k=4

doc d - a rose is a rose is a rose.

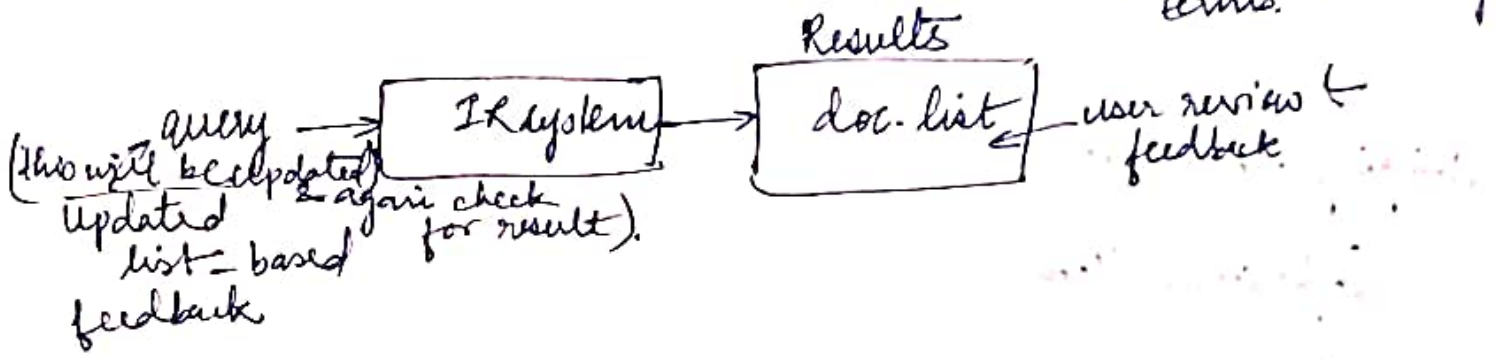


1) a rose is a
2) rose is a rose
3) is a rose is
4) a rose is a
5) rose is a rose

08/10/24

Query Expansion – Technique to increase the recall.
|
which retrieves the relevant data

(To improve the query-storing so that more relevance doc. are returned)
Relevance feedback – Add more terms / Delete unnecessary terms.



JACCARD CO-EFFICIENT

$$J(A,B) = \frac{A \cap B}{A \cup B}$$

$$J[S(d_1), S(d_2)] = \frac{S(d_1) \cap S(d_2)}{S(d_1) \cup S(d_2)} \geq \kappa$$

d_1 & d_2 are similar

can find out the similarity

$K = 4$.

$d_1 \longrightarrow$ This is a test for IR
$d_2 \longrightarrow$ This is a test for IR Assessments
threshold = 0.7

Shingling ←

$S(d_1)$	$S(d_2)$
① This is a test	① This is a test
② is a test for	② is a test for
③ a test for IR	③ a test for IR
	④ test for IR Assessments

08/10/24

$J(S(d_1), S(d_2)) = \frac{3}{4} = 0.75 > th_r(0.7)$

$\therefore$ d_1 & d_2 are similar.

Q.

$S(d_1)$	$S(d_2)$	$S(d_3)$
1	0	1
1	1	0
1	1	1
1	1	1
1	1	0

identify which docs are similar if threshold for $J.C = 0.5$.

$J(S(d_1), S(d_2)) = \frac{4}{5} = 0.8 > th.$ — similar

$J(S(d_2), S(d_3)) = \frac{2}{5} = 0.4 < th.$ — Not similar

$J(S(d_3), S(d_1)) = \frac{3}{5} = 0.6 \geq th$ — similar

Q.

$S(d_1)$	$S(d_2)$	$S(d_3)$	$S(d_4)$
1	0	1	1
1	1	0	0
1	1	1	1
1	1	1	1
1	1	0	1

$th = 0.8$

06.10.24

$^4C_2 = \frac{4!}{2!\,2!} = \frac{1\times2\times3\times4}{2\times1\times2\times1} = 6$

~~J(d1, d3)~~

U

d_i	d_j	$J(S(d_i), S(d_j))$
d_1	d_2	$4/5 = 0.8 \geq th$ ✓
d_2	d_3	$2/5 = 0.4 < th$ ✗
d_3	d_4	$3/4 = 0.75 \leq th$ ✗
d_4	d_1	$4/5 = 0.8 \geq th$ ✓
d_1	d_3	$3/5 = 0.6 \leq th$ ✗
~~d2~~	~~d4~~	
d_2	d_4	$3/5 = 0.6 < th$ ✗

d_1 & d_2 } similar.
d_4 & d_1 }