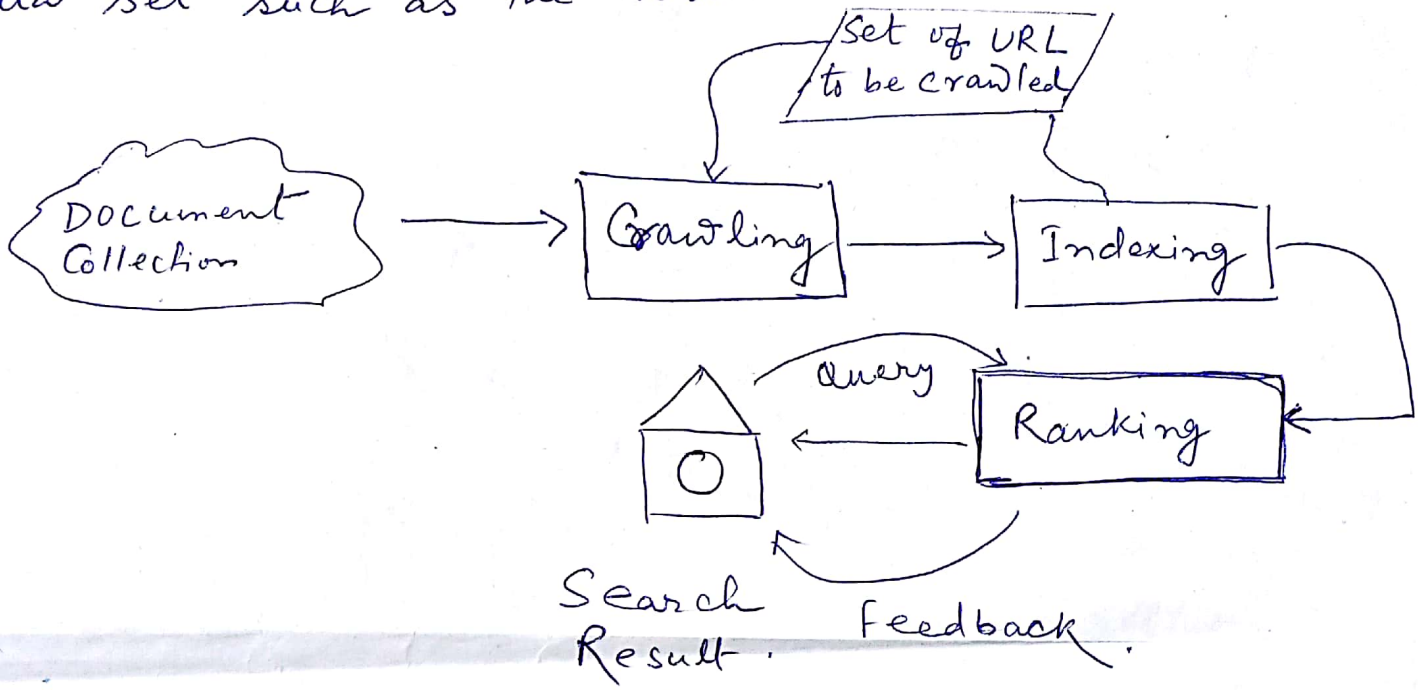
Web Based Information Retrieval

: Searching for relevant documents and information within the contents of specific data set such as the WWW.



Crawling : The system browses the document collection and fetches documents.

Indexing : The system builds an index of the documents fetched during crawling.

Ranking : The system retrieves documents that are relevent to the user

Feedback : The initial result returned from a given query may be used to refine the query.

Page Rank Algorithm: Page Rank is one of the google algorithm used by search engine to rank websites in their search engine result page. It was developed by Larry Page and Sergey Brin to rank webpages in search engine.

This is the measure of incoming links given by all other pages on the web to know about the importance of that particular page.

In brief the algorithm works as follows:

$$PR(A) = (1-d) + d(PR(T_1)/C(T_1)) + \cdots + PR(T_n)/C(T_n))$$

where,

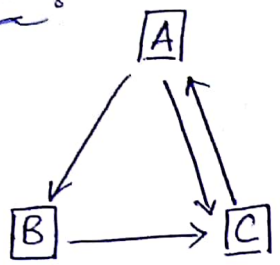
- $PR(A)$ => the pageRank of page A
- $PR(T_i)$ => the pageRank of page T_i which link to page A.
- $C(T_i)$ => the number of outbound links on page T_i and
- d => the damping factor, which can be set between 0 and 1.

Formally, * PageRank is an iterative algorithm. The PageRank value change in each step.

PageRank of a given page (Webpage) = Initial PageRank + (total ranking power + number of outbound links) + ...

NOTE: For a random surfer model the minimum page rank = $(1-d)$ and maximum PageRank = $N + (1-d)$, where N is the total number of webpages linked each other.

Example :



Consider an imaginary web, containing 3 web pages.

Also, the inbound and outbound link structure is shown in the figure.

Also, let the damping factor $d = 0.5$.

Find the pageRanks of A, B and C.

Sol.

Now, the calculation of PageRanks of A, B, and C can be done using following relations.

$$PR(A) = 0.5 + 0.5\,PR(C) \quad \text{—(i)}$$

$$PR(B) = 0.5 + 0.5\,(PR(A)/2) \quad \text{—(ii)}$$

$$PR(C) = 0.5 + 0.5\,(PR(A)/2 + PR(B)) \quad \text{—(iii)}$$

NOTE : $d = 0.5$
$\Rightarrow 1 - d = 0.5$

Solving equ^{n} (i) (ii) & (iii) iteratively

$PR(A) = 0.5 + (0.5 * 1)$ ~~$= 0.5 + 0.5\,(PR(A)/2 + PR(B))$~~

$= 1$

Consider the initial pageRank $PR(C) = 1$

~~$\Rightarrow PR(A) = 1/4 + 1/4\,(1/2)$~~

$PR(B) = 0.5 + 0.5\,(1/2)$

$= 0.5 + 0.25$

$= 0.75$

$PR(C) = 0.5 + 0.5\,((PR(A)/2) + PR(B))$

$= 0.5 + 0.5\,(1/2 + 0.75)$

$= 0.5 + 0.625$

$= 1.125$