Date _____ Day _____

"PAGE RANK ALGORITHM"

→ The pagerank of any page is roughly the probability that the random surfer will land on a particular page.

→ The behaviour of the google's random surfer is an example of a Markov process, that depends only of the current state of a system and not on its history.

→ Formula:

$$PR(p) = \frac{(1-d)}{N} + d \sum \frac{PR(q)}{L(q)}$$

summation of incoming links to page p

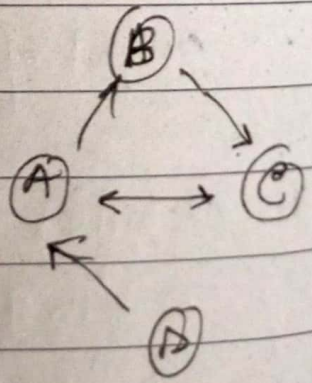
L(q) → page q k no. of outlinks

d = 0.85

N = Total no. of pages

Initial Page rank will be 1/N for all pages.

We perform iterations until our values converges.



Date _____ Day _____

$N = 4$, $d = 0.85$ $\Rightarrow (1 - 0.85)/4 = 0.0375$

→ Iteration #01

$$P(A) = 0.0375 + 0.85\left(\frac{PR(C)}{1} + \frac{PR(D)}{1}\right)$$

$$= 0.0375 + 0.85(0.25 + 0.25)$$

$$= 0.4625$$

$$P(B) = 0.0375 + 0.85[PR(A)/2)$$

$$= 0.14375$$

$$P(C) = 0.0375 + 0.85(P(A)/2 + P(B)/1)$$

$$= 0.35625$$

$$P(C) = 0.0375 + 0.85(0)$$

$$= 0.0375$$

→ Iteration #02

$PR(A) = 0.37219$

$PR(B) = 0.23406 \Rightarrow 0.0375 + 0.85(0.4625/2)$

$PR(C) = 0.35625 \Rightarrow 0.0375 + 0.85(0.4625/2 + 0.14375)$

$PR(D) = 0.0375$

Date _____ Day _____

"HYPERLOGLOG ALGO"

→ This algo is used to count "No. of distinct users". (approximate ans not accurate).

→ Let suppose we have usernames stored..

→ Each username is converted to a binary string using some Hash Function.

ankush → 001100

samroze → 010010

ali → 011011

shaheer → 000101

⋮

- Example
  - 1001
  - 0000
  - 1100
  - 1100
  - 0011

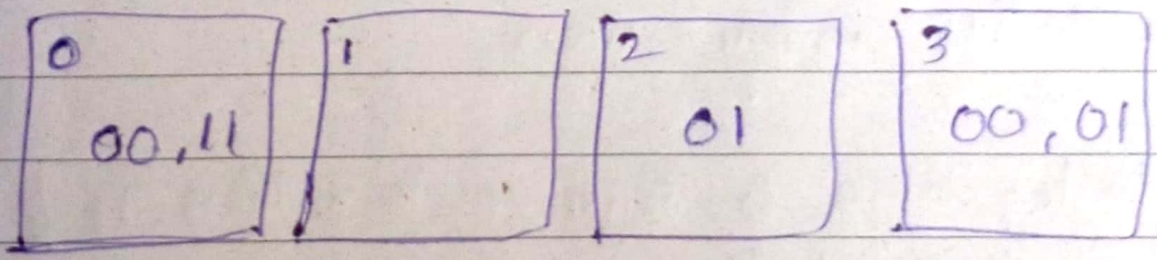
We have "m" buckets we need to place these strings in the respective buckets.

Page No. ____

Date _____ Day _____

→ Let say we have m = 4.
→ Index of buckets → 0, 1, 2, 3
→ The 1st 2 bits of the string represents the bucket index, and the remaining bits of the string will be stored in bucket.

10 01 → bucket 2
00 00 → bucket 0
11 00 → bucket 3
11 01 → bucket 3
00 11 → bucket 0



→ Ab har ek bucket mein max no. of consecutive zeros from left check krenge.

→ $b_0 = 2$, $b_1 = 0$, $b_2 = 1$, $b_3 = 2$

Date LogLog Formula: Day

Now take mean,

$$\frac{2+0+1+2}{4} = 1.25$$

$$= \text{constant} * m * 2^{a_m}$$

where,

Constant = 0.79

m = no. of buckets = 4

a_m = mean

$$= 0.79 \times 4 * 2^{1.25}$$

$$= 7.515 \text{ unique users}$$

→ but loglog algo is sensitive to outliers Therefore we use hyperloglog.

HyperLogLog Formula:

→ instead of simple mean we take harmonic me

$$= \text{constant} * m * \left(m / \sum_{i=1}^{m} \frac{1}{2^{\text{bucket}[i]}} \right)$$

Date _____ Day _____

$$= 0.79 \times 4 \times \frac{4}{1/2^2 + 1/2^0 + 1/2^1 + 1/2^2}$$

$$= 6.32$$