Need: Informational (Google)

navigational (Blackboard). (following a linear path).

transaction (online shop). (maybe some stages).

(or sharing materials resources)

Problems: volatile data: how to fresh the index or the content of the index.

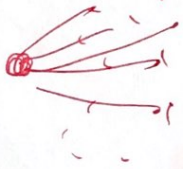
scaling issues.

redundancy: how to identify it is similar. (diversity)

quality ~~relevance or~~

diversity: rank on relevant.

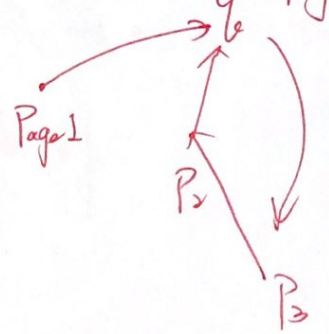
rank on space (which material is the most important?



Evaluation of web search:

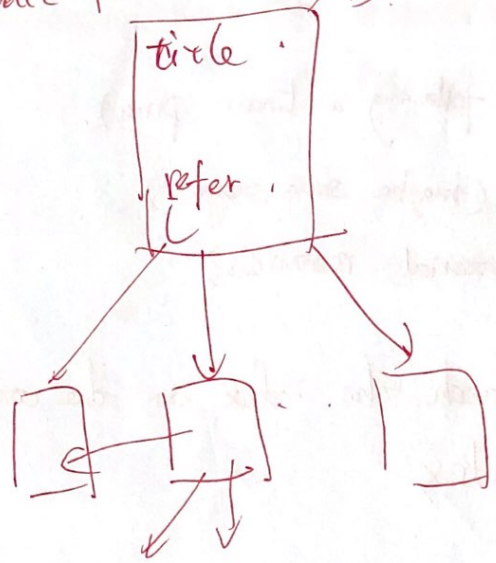
1st generation: tf-idf → no way to guarantee the return results are relevant. (Page q).

2nd generation: Google. link analysis



3rd generation:

Impact factor = A/B.



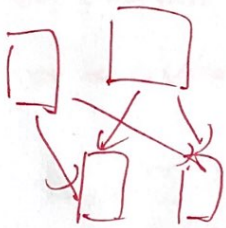
directive graph

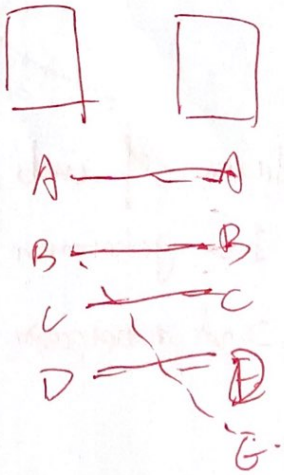
cycle free.

Sometime can weight the new but important papers a little bit.

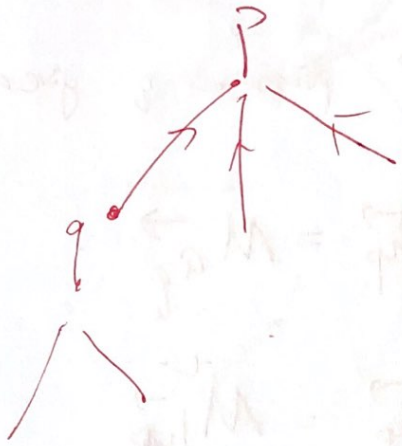
~~similar analysis. co-citation~~

another measure of similarities : co-citation





Mutually recursive heuristics used (authority. hub).



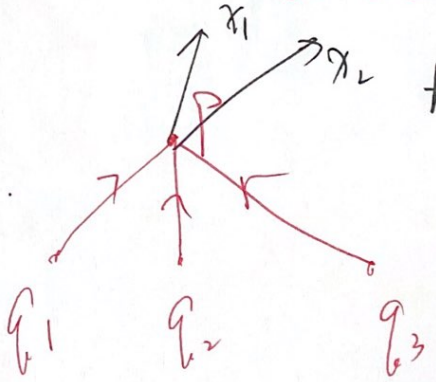
increase the rank.

$\vec{a_p}$: p_1 p_2 p_3 ... $\in S$

$\vec{h_p}$: p_1 p_2 p_3 ... $\in S$.

$$a_p = \sum_{q : q \to p} h_q .$$

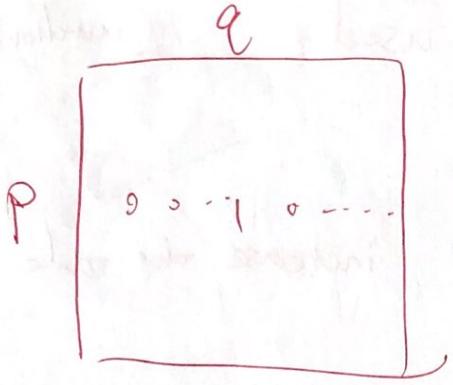
initially set every score to constant. don't need to recalculate.



$$h_p = \sum_{q : p \to q} a_q$$

We want to return the $\vec{a_p}$: p_1 p_2 p_3 $\to \in S$. to the user with the order of

Metrix



"p" points to "q" give a value=1

$$\vec{h_p} = M \vec{a_q}$$

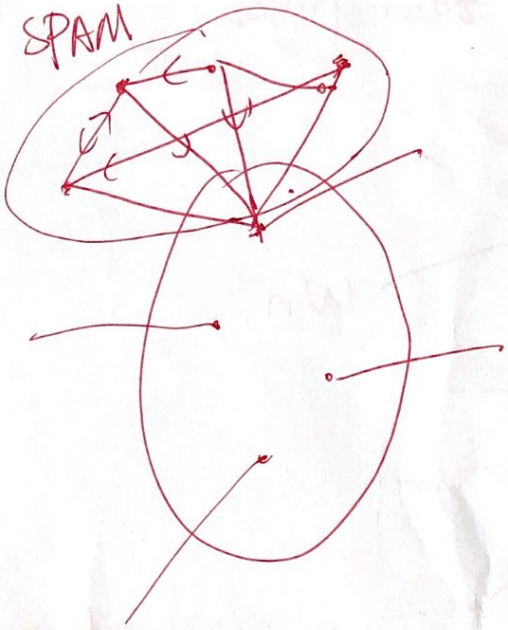
$$\vec{a_p} = M^t \vec{h_p}$$

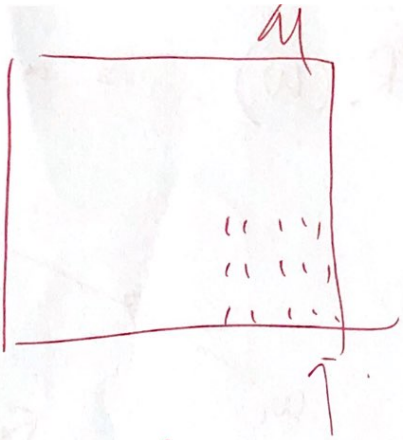
(transpose metrics)

$$\Rightarrow \vec{a_p} = (M^t M)\vec{a_q}$$

(purely function of metrics)

give the ranking

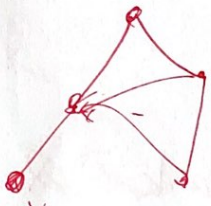




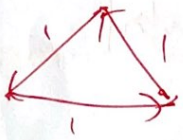
the more edges goes here
the ~~more~~ less likely go promoted.

Page Rank

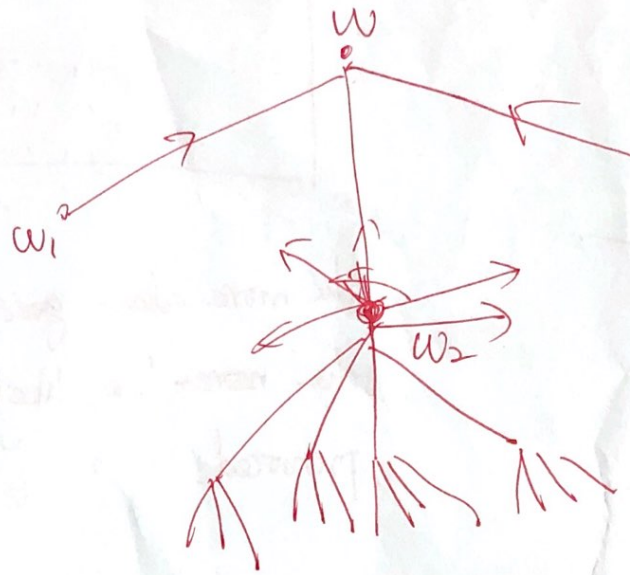
Problem.



no where to go ⇒. means have equally small prob to every page.

page sink  then it empty.

PR(w). is a function of other pages.



teleportation probability $\frac{1}{N} \times (1-T)$.

↑ times happen.

↓ the number of ——.

Difference between PageRank & HITS.

PageRank: query independent. global.

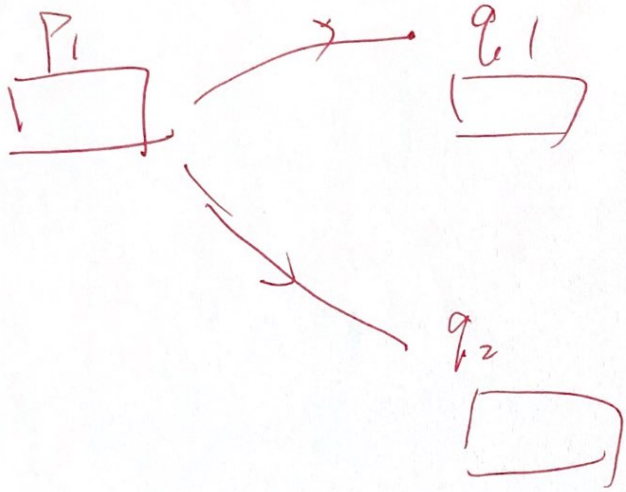
HITS: ↑ query specific. local graph.

distribute the web. based on user's interests.

0 0 3 1 2 .

P_1 P_2 P_3 ... P_n.

$\frac{1}{N}$ $\frac{1}{N}$ $\frac{1}{N}$ $\frac{1}{N}$.



$sim(P_1, q_1)$

$sim(P_1, q_2)$.

Rank signals:

content signals: $sim(q, d)$

structural signals: $sim_c(q, d)$.

web usage: PR.

link based ranking PR.

$R(P, q) = (\alpha)$ BM25 $+ (1-\alpha)$ . . ( )

↓ how importance the query is

↳ the query is hard. we can put more weight on BM25 ( ).