Search engines: ranking algorithms

Gianna M. Del Corso

Dipartimento di Informatica, Università di Pisa, Italy

LSMC, 1 Aprile 2014

1. Statistics

2. Search Engines
    - Ranking Algorithms
    - HITS
    - PageRank

Web Analytics

- Estimated size: up to 45 billion pages
- Indexed pages: at least 2 billion pages
- Average size of a web page is 1.3 MB.... Up 35% in last year
- Petabytes (10^{15} bytes) for storing just the indexed pages!!

Web Analytics

- Web pages are dynamical objects
- Most of web pages content changes daily
- Average lifespan of a page is around 10 days.

A huge quantity of data!!

Find a document without search engines it's like looking for a needle in a haystack.

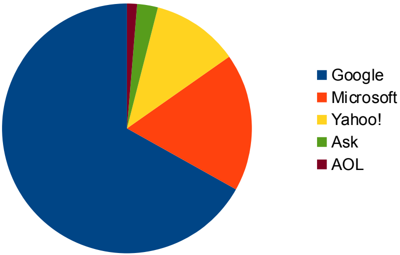
Statistics

Market shares

- https://www.quantcast.com/top-sites

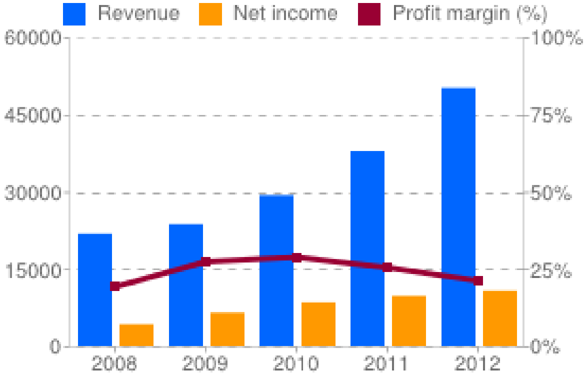
Statistics

Market shares

- https://www.quantcast.com/top-sites
- 

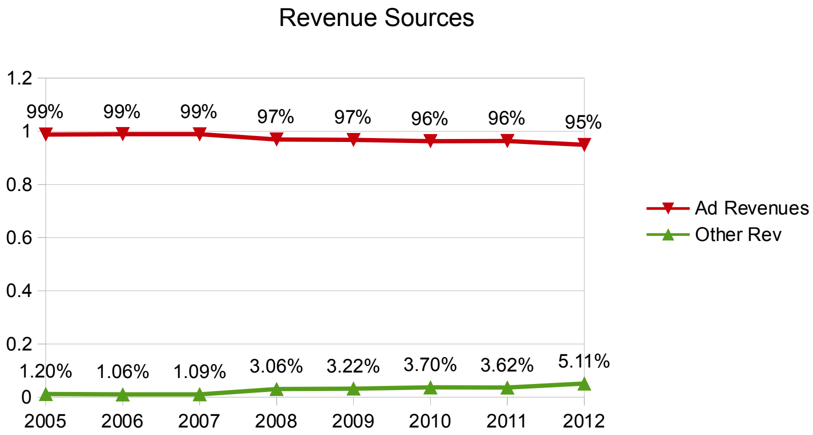
[comScore, July 2013]

Google's Numbers



[NASDAQ:GOOG]

Google's Numbers

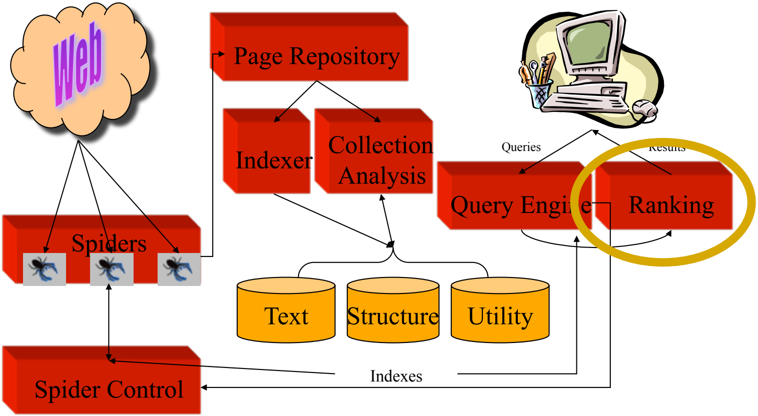


[investor.google.com/financial/tables.html]

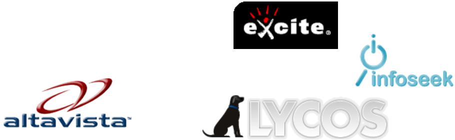
Search Engines

- The dream of search engines: catalog everything published on the web
- Web content fruition by query
- Rank by relevance to the query

The Anatomy of a Search Engine



... At the very beginning (mid ’90s)



The order of pages returned by the engine for a given query was controlled by the owner of the page, who could increase the ranking score increasing the frequency of certain terms, font color, font size etc.

SPAM

… At the very beginning (mid '90s)

The order of pages returned by the engine for a given query was controlled by the owner of the page, who could increase the ranking score increasing the frequency of certain terms, font color, font size etc.

SPAM

Modern engines

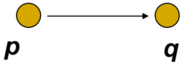
In 1998 two similar ideas

- HITS (Jon Kleinberg)
- PageRank (S. Brin and L. Page)

The importance of a page should not depend on the owner of the page!

Basic intuition

Look at the linkage structure



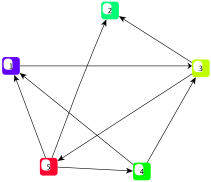
Adding a link from page p to page q is a sort of vote from the author of p to q.

Idea If the content of a page is interesting it will receive many "votes", i.e. it will have many inlinks.

Web Graph

The Web is seen as a directed graph:

- Each page is a node
- Each hyperlink is an edge



$$G = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \end{bmatrix}$$

Web Graph

- The importance of a web page is determined by the structure of the web graph
- At first approximation: Contents of pages is not used!
- Aim: The owner of a page cannot boost the importance of its pages!

An example

query=the best automobile makers in the last 4 years

Intention get back a list of top car brands and their official web sites

Textual ranking pages returned might be very different

Top companies might not even use the terms "automobile makers".They might use the term "car manufacturer" instead

HITS

Each page has associated two scores

a_i authority score

h_i hub score

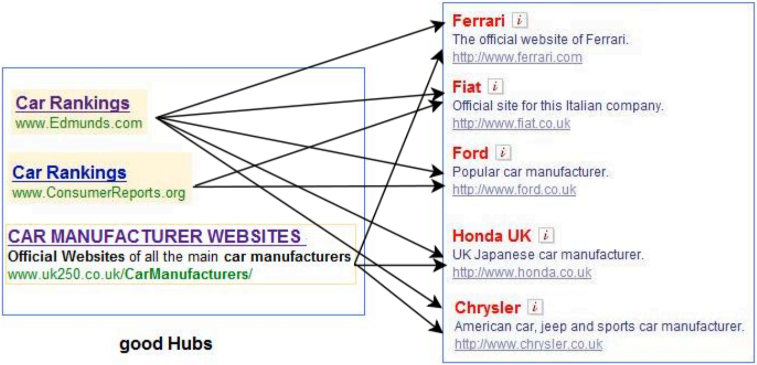
A page is a good “authority” if it is linked by many good hubs
A page is a good “hub” if it is linked by many good authorities
To every page p we associate two scores:

$$\begin{cases} a_p & = \sum_{i \in \mathcal{I}(p)} h_i \\ h_p & = \sum_{i \in \mathcal{O}(p)} a_i \end{cases}$$

$$\begin{cases} \mathbf{a}^{(i)} & = G^T \mathbf{h}^{(i-1)} \\ \mathbf{h}^{(i)} & = G \mathbf{a}^{(i)} \end{cases}$$

Hub pages advertise authoritative pages!

HITS



Car Rankings
www.Edmunds.com

Car Rankings
www.ConsumerReports.org

CAR MANUFACTURER WEBSITES
Official Websites of all the main **car manufacturers**
www.uk250.co.uk/**CarManufacturers/**

Ferrari i
The official website of Ferrari.
http://www.ferrari.com

Fiat i
Official site for this Italian company.
http://www.fiat.co.uk

Ford i
Popular car manufacturer.
http://www.ford.co.uk

Honda UK i
UK Japanese car manufacturer.
http://www.honda.co.uk

Chrysler i
American car, jeep and sports car manufacturer.
http://www.chrysler.co.uk

good Hubs

good Authorities

Query: Top automobile makers

HITS

- G^TG and GG^T are non negative
- G^TG and GG^T are semipositive defined
- real nonnegative eigenvalues

$$\begin{cases} \mathbf{h}^{(i)} & = GG^T\mathbf{h}^{(i-1)} \\ \mathbf{a}^{(i)} & = G^TG\mathbf{a}^{(i-1)} \end{cases}$$

$\mathbf{h}^*$ is the dominant eigenvector of GG^T
$\mathbf{a}^*$ is the dominant eigenvector of G^TG

For Perron-Frobenius Theorem the eigenvector associated to the dominant eigenvalue has nonnegative entries!

HITS

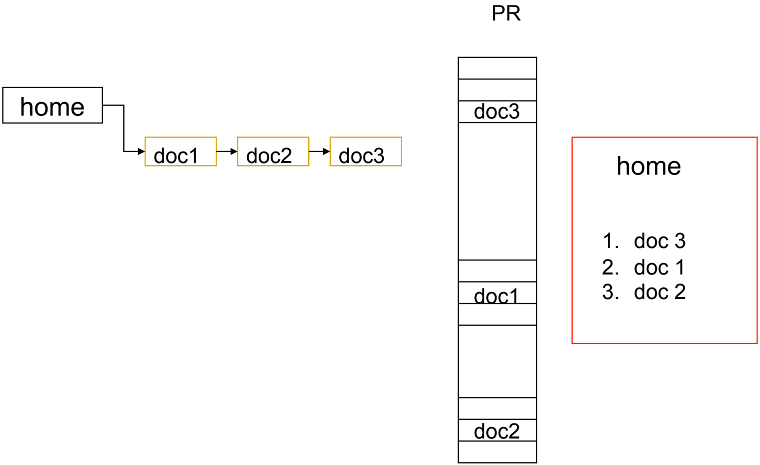
At query time

- Find relevant pages
- Construct the graph starting with those bunch of pages
- Compute dominant eigenvector of $G^T G$
- Sort the dominant eigenvalue and return the pages in accordance with the ordering induced

Google's PageRank

- Is a static ranking schema
- At query time relevant pages are retrievedneighbours
- The ranking of pages is based on the PageRank of pages which is precomputed

PageRank



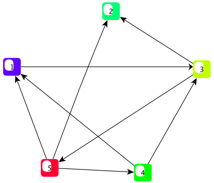
PageRank

- A page is important if is voted by important pages
- The vote is expressed by a link

PageRank

- A page distribute its importance equally to its neighbors
- The importance of a page is the sum of the importances of pages which points to it

$$\pi_j = \sum_{i \in \mathcal{I}(j)} \frac{\pi_i}{\mathrm{outdeg}(i)}$$



$$G = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \end{bmatrix} P = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 1/2 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 1/3 & 1/3 & 0 & 1/3 & 0 \end{bmatrix}$$

PageRank

It is called Random surfer model

The web surfer jumps from page to page following hyperlinks. The probability of jumping to a page depends of the number of links in that page.

Starting with a vector $\mathbf{z}^{(0)}$, compute

$$\mathbf{z}_j^{(k)} = \sum_{i \in \mathcal{I}(j)} \mathbf{z}^{(k-1)} p_{ij}, \quad p_{ij} = \frac{1}{\text{outdeg}(i)}$$

Equivalent to compute the stationary distribution of the Markov chain with transition matrix P.

Equivalent to compute the left eigenvector of P corresponding to eigenvalue 1.

PageRank

It is called Random surfer model

The web surfer jumps from page to page following hyperlinks. The probability of jumping to a page depends of the number of links in that page.

Starting with a vector $\mathbf{z}^{(0)}$, compute

$$\mathbf{z}_j^{(k)} = \sum_{i \in \mathcal{I}(j)} \mathbf{z}^{(k-1)} p_{ij}, \quad p_{ij} = \frac{1}{\mathrm{outdeg}(i)}$$

Equivalent to compute the stationary distribution of the Markov chain with transition matrix P.

Equivalent to compute the left eigenvector of P corresponding to eigenvalue 1.

PageRank

It is called Random surfer model

The web surfer jumps from page to page following hyperlinks. The probability of jumping to a page depends of the number of links in that page.

Starting with a vector $\mathbf{z}^{(0)}$, compute

$$\mathbf{z}_j^{(k)} = \sum_{i \in \mathcal{I}(j)} \mathbf{z}^{(k-1)} p_{ij}, \quad p_{ij} = \frac{1}{\mathrm{outdeg}(i)}$$

Equivalent to compute the stationary distribution of the Markov chain with transition matrix P.

Equivalent to compute the left eigenvector of P corresponding to eigenvalue 1.

PageRank

Two problems:

- Presence of dangling nodes
  - P cannot be stochastic
  - P might not have the eigenvalue 1
- Presence of cycles
  - P can be reducible: the random surfer is trapped
  - Uniqueness of eigenvalue 1 not guaranteed

Perron-Frobenius Theorem

Let $A \geq 0$ be an irreducible matrix

- there exists an eigenvector equal to the spectral radius of A, with algebraic multiplicity 1
- there exists an eigenvector $\mathbf{x} > \mathbf{0}$ such that $A\mathbf{x} = \rho(A)\mathbf{x}$.
- if $A > 0$, then $\rho(A)$ is the unique eigenvalue with maximum modulo.

PageRank

Presence of dangling nodes

$$\bar{P} = P + \mathbf{d}\mathbf{v}^T$$

$$d_i = \begin{cases} 1 & \text{if page } i \text{ is dangling} \\ 0 & \text{otherwise} \end{cases} \qquad v_i = 1/n;$$

$$P = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 1/2 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 1/3 & 1/3 & 0 & 1/3 & 0 \end{bmatrix} \bar{P} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 0 & 1/2 & 0 & 0 & 1/2 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 1/3 & 1/3 & 0 & 1/3 & 0 \end{bmatrix}$$

PageRank

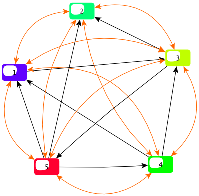
Presence of cycles
Force irreducibility by adding artificial arcs chosen by the random surfer with "small probability" α.

$$\bar{\bar{P}} = (1-\alpha)\bar{P} + \alpha \mathbf{e}\mathbf{v}^T,$$

$$\bar{\bar{P}} = (1-\alpha)\begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 0 & 1/2 & 0 & 0 & 1/2 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 1/3 & 1/3 & 0 & 1/3 & 0 \end{bmatrix} + \alpha \begin{bmatrix} 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \end{bmatrix}.$$

Typical values of α is 0.15.

PageRank: a toy eample



$$\bar{\bar{P}} = \begin{bmatrix} 0.03 & 0.03 & 0.88 & 0.03 & 0.03 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.03 & 0.45 & 0.03 & 0.03 & 0.45 \\ 0.45 & 0.03 & 0.45 & 0.03 & 0.03 \\ 0.31 & 0.31 & 0.03 & 0.31 & 0.03 \end{bmatrix}$$

Computing the largest eigenvector of $\bar{\bar{P}}$ we get

$$\mathbf{z}^T \approx [0.38, 0.52, 0.59, 0.27, 0.40],$$

which corresponds to the following order of importance of pages

$$[3, 2, 5, 1, 4].$$

PageRank

- P is sparse, $\bar{\bar{P}}$ is full.
- The vector $\mathbf{y} = \bar{\bar{P}}^T\mathbf{x}$, for $\mathbf{x} \geq 0$, such that $\|\mathbf{x}\|_1 = 1$ can be computed as follows

$$\begin{aligned} \mathbf{y} &= (1-\alpha)P^T\mathbf{x} \\ \gamma &= \|\mathbf{x}\| - \|\mathbf{y}\|, \\ \mathbf{y} &= \mathbf{y} + \gamma\mathbf{v}. \end{aligned}$$

- The eigenvalues of $\bar{P}$ and $\bar{\bar{P}}$ are interrelated:

$$\lambda_1(\bar{P}) = \lambda_1(\bar{\bar{P}}) = 1, \quad \lambda_j(\bar{\bar{P}}) = (1-\alpha)\,\lambda_j(\bar{P}), j > 1.$$

- For the web graph $\lambda_2(\bar{\bar{P}}) = (1-\alpha)$.