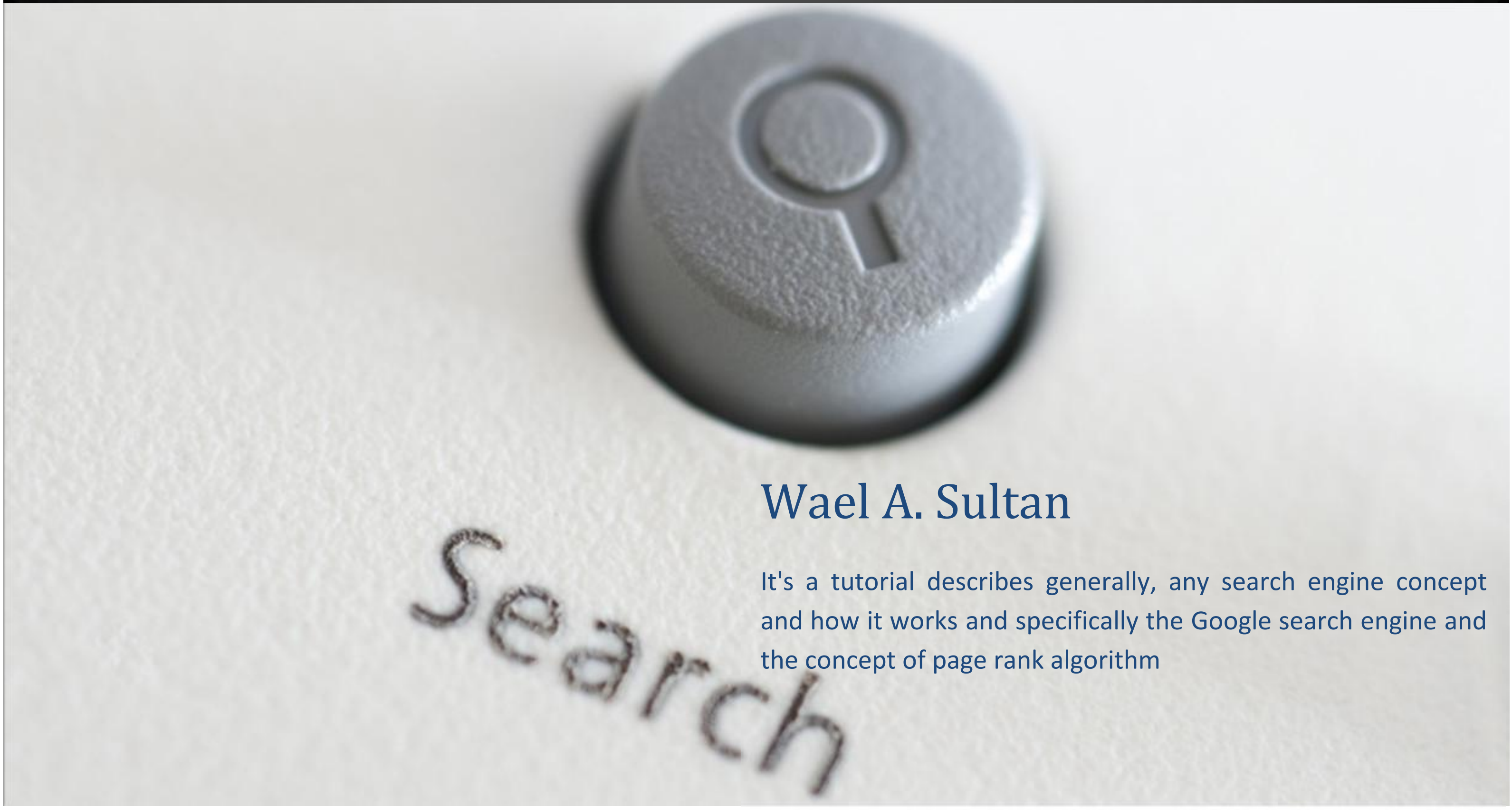
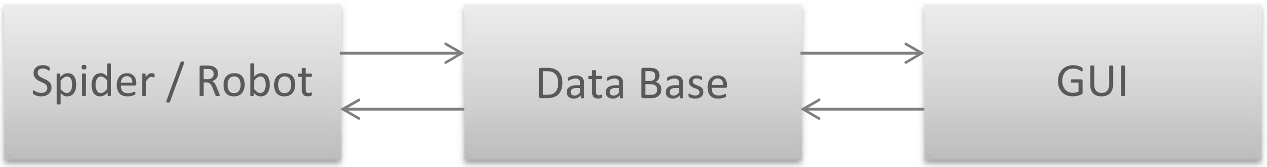
Search Engine and PageRank



Wael A. Sultan

It's a tutorial describes generally, any search engine concept and how it works and specifically the Google search engine and the concept of page rank algorithm

Block Diagram of any web search



Spider / Robot

Data Base

GUI

Spider: Take the human out of the equation by automatically crawling the web and constantly searching for texts, pictures… etc. and it copies everything back to the search engine database.

Briefly, Spider is a search engine of the search engine

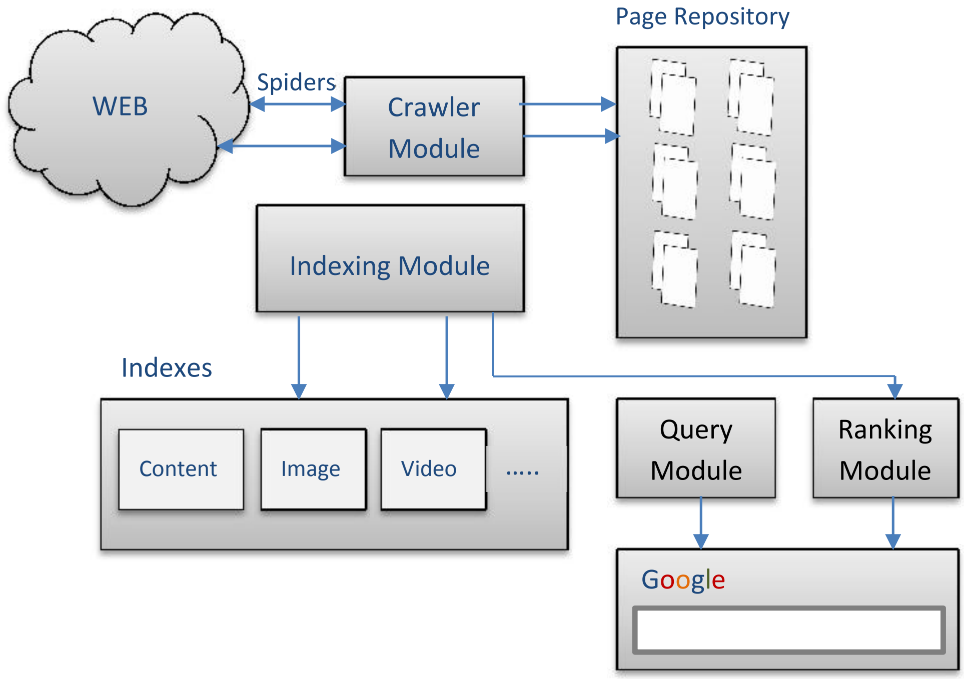
Search engine database stores and organize everything the spider give to it. And it simply is thousands of thousands of servers.

Specifically, Google has about 100,000 servers as a database.

What we see in the web site

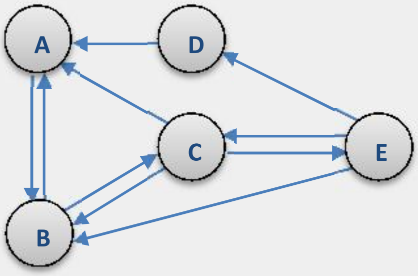
Remember: when you search in any web search engine you not search the internet, you search copy of internet that is pre-stored in that search engine database

Block Diagram Google search engine



Google Search Engine and PageRank

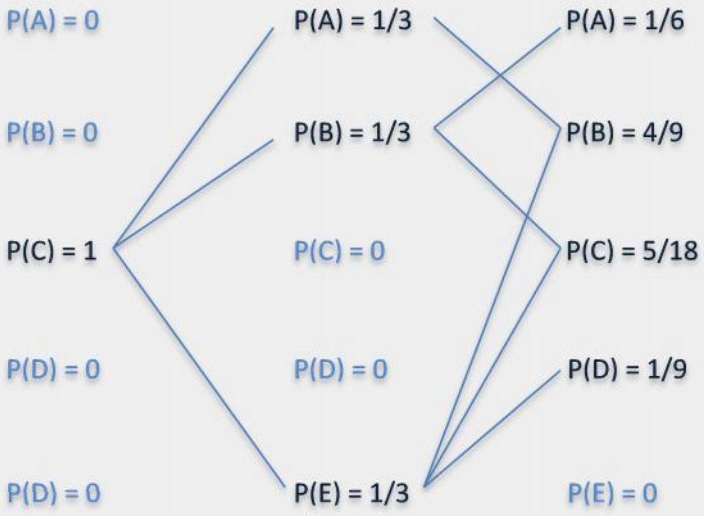
The web is composed of billions of individual Pages, and more links between them. As such, the web can be modeled as a directed graph where pages are nodes, and links are directed edges between them. We suggest a simple case and then generalize the concept.



In this small web containing five pages (A, B, C, D, and E).
The directed edges between the nodes indicate that

- the only link from page A leads to page B,
- page B links to pages A and C,
- page C links to pages A, B, and E,
- the only link from page D leads to page A, and
- page E links to pages B, C, and D.

In order to determine the ranking to be accorded to each of these five pages , we consider a simple version of the PageRank algorithm.
Suppose that an impartial web surfer navigates through this web by randomly choosing links to follow.
Consider the imperial web surfer starting at page C the following figure represent the probable path and corresponding conditional probabilities.



We can deduce the first step probabilities table as following and we called it the transition matrix P for simple web above

$$p = \left(\begin{array}{cccccc} A & B & C & D & E & {}^{Nodes}/_{Nodes} \\ 0 & \frac{1}{2} & \frac{1}{3} & 1 & 0 & A \\ 1 & 0 & \frac{1}{3} & 0 & \frac{1}{3} & B \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{3} & C \\ 0 & 0 & 0 & 0 & \frac{1}{3} & D \\ 0 & 0 & \frac{1}{3} & 0 & 0 & E \end{array}\right)$$

Now and as before, we assume that the web crawler starts at page C. Thus

$$P^0 = \begin{pmatrix} p(X_0 = A) \\ p(X_0 = B) \\ p(X_0 = C) \\ p(X_0 = D) \\ p(X_0 = E) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

The probability vector p^1 after the first step is given by $p^1 = P\,p^0$,and therefore

$$P^1 = \begin{pmatrix} p(X_1 = A) \\ p(X_1 = B) \\ p(X_1 = C) \\ p(X_1 = D) \\ p(X_1 = E) \end{pmatrix} = p = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{3} & 1 & 0 \\ 1 & 0 & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 0 & \frac{1}{3} \\ 0 & 0 & \frac{1}{3} & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{3} \\ \frac{1}{3} \\ 0 \\ 0 \\ \frac{1}{3} \end{pmatrix}$$

The same method may be followed to calculate the probability vector after any number of steps:

$$p^n = P\,p^{n-1} = PP\,p^{n-2} = \cdots = P^n\,p^0$$

We observe that seems to converge to a constant matrix as m increases, in our case:

$$p^{32} = \begin{pmatrix} 0.293 & 0.293 & 0.293 & 0.293 & 0.293 \\ 0.390 & 0.390 & 0.390 & 0.390 & 0.390 \\ 0.220 & 0.220 & 0.220 & 0.220 & 0.220 \\ 0.024 & 0.024 & 0.024 & 0.024 & 0.024 \\ 0.073 & 0.073 & 0.073 & 0.073 & 0.073 \end{pmatrix}$$