Calculating Page Pertinence for Web Referencing a New Approach

H. Barboucha
Labo Matsi, ESTO, B.P 473,
University Mohammed I OUJDA, Maroc

M. Nasri
Labo Matsi, ESTO, B.P 473,
University Mohammed I Oujda,
Maroc

ABSTRACT

Every day, more than three billion queries are performed on Google [1]. To take advantage of opportunities that you're offered by the search engines, it is important for your website to achieve high rankings in search results, in order to reach a larger customer base, increasing its business in other regions (or countries). One goal of any business is to convert a maximum of Internet users into potential customers. ALGORITHM[2] links that allowed Google to be so successful and become what it is today! But how algorithm function to rank sites in terms of importance ? What formulas are the basis for website ranking ? The developers carry out tests and then draw conclusions. The purpose of this paper is to find a method (algorithm) which calculates the relevance of each page, we unveil parts of the code of the largest search engine in the world.

Keywords

E-commerce, algorithm, SEO web, social networks, reputation, SEO, SMO, page rank, HTML5, CSS3, backlink, in page, off page. googlebot

1. INTRODUCTION

On the web, links are synonymous with quality. If a site A has a link to a site B, it is because site B offers relevant content ... and if site B itself establishes a link to a site C, it is because site C is quite enough interesting to site B, which is already recognized as relevant by Site A and thus make a link to it . And the same applies to site D. GOOGLE has translated this idea in a page indexing algorithm that it still uses today, and which has allowed it to emerge as the search engine reference . The web is nothing but a huge pile of websites that exist next to each other. Sometimes these sites are partners, some other times they are competitors.

Nevertheless, all these sites want the one same thing: to be visited.

Web links that are so performing have allowed Google to become so powerful! But how do such links that help rank sites in order of importance function? To answer this question, we conducted tests on three different websites.

The first part is an introduction to web links and their functioning. The second part of this work is devoted to the different structures of a site and the methods of calculating links. The third part describes the link calculation iterations. In conclusion, we are making recommendations as to the good structure and the appropriate practices that should satisfy all webmasters and marketing managers who wish to update their knowledge on SEO.

2. CALCULATION OF THE IMPORTANCE OF LINKS

In the real world web sites have a large number of pages with multiple links and different structures.
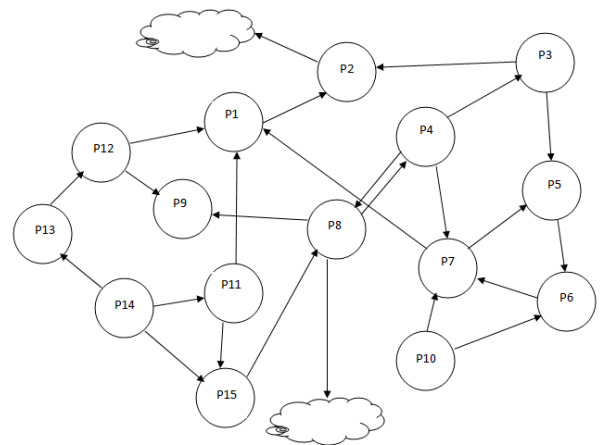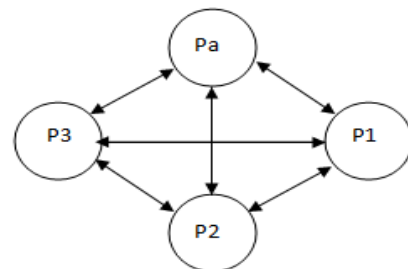

Fig1: structure of links between pages of a website.

The links[3] on the internet are not random but have been carefully edited. What information could this graph give us?
To analyze the structure of a website we will neglect the content of pages and only consider the links between them.
In this study we selected three different structures, simplified for the sake of simplification of the calculations.

Site 1: a site has four pages each of which sends links to all other pages
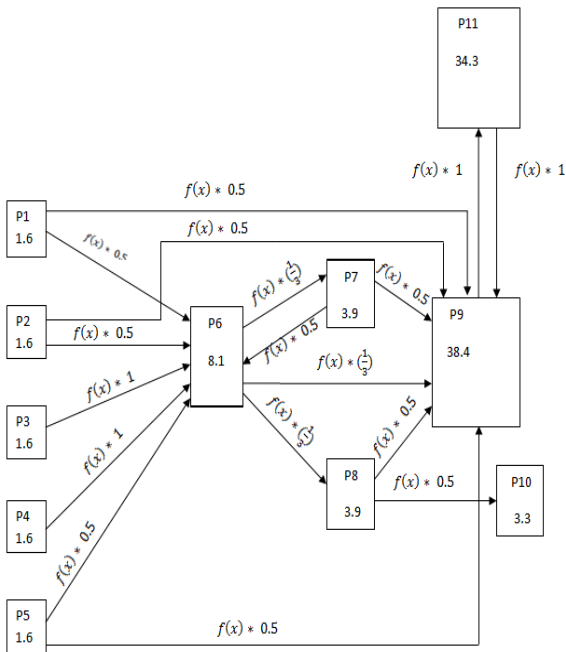
Site 2: In this site, the home page (Pa) sends a link to the site map (P1). The latter, in addition to the return to the home page link sends a link to each page of the site (p2, p3, p4), these pages emit a link back to the homepage

Site 3: here the homepage (Pa) sends a link to each of the category pages (P1 P2 P3), the categories send back links to the home page in addition to sending links to other pages on the site

To calculate the relevance[4] or importance of the pages of each website it is necessary to calculate the importance of the incoming links to each page.

Links on a web page that allowed Google to be so successful and become what it is today! But how do these links that classifies sites by importance work?

The basic idea, yet to formalize, is that a link j -> i is a recommendation of page Pj to go read the page Pi. This is thus a Pj vote in favor of the authority of the page Pi, just like the principle of elections.
Comparison with the election stops there because all the pages do not have the same power to "vote".

According to the tests we carried out on three different websites (diagram 1, 2,3), we noted that the ranking of websites is based on the application of the following formulas:

2.1 Calculation of incoming links (in-page)

The first thing to notice is that an important page receives many links. With a little naivety, we would assume the reciprocal affirmation: if a page gets a lot of links, so it is important. Thus we could define the importance f(x) of page x as the number of page links yi toward page x. As a formula, it is written as follows:

$$f(x) = \sum yi \to x$$

2.2 Outgoing links (off-page)

Some pages emit a lot of links: they seem less specific and their weight will be lower. We therefore agree with the vote of the page yi as L(yi) equally, where L (yi) denotes the number of connections issued by page yi. Thus we could define a more accurate measurement:

$$f(x) = \sum_{yi \to x} \frac{1}{L(yi)}$$

2.3 Relevance of source pages

A Page x looks important if many important pages are mentioning it. This leads us to define the importance f (x) in a recursive way, as follows:

$$f(x) = \sum_{yi \to x} \frac{f(yi)}{L(yi)}$$

Here the weight of the vote yi for x is proportional to the weight f(yi) of the issuing page. It is easy to state but less obvious to calculate.

2.4 The damping factor

The idea of browsing the web is that a user (imaginary) who is randomly clicking on links will continue to click and come across a specific page, this probability is the damping factor[5] noted (d).

Fig 1: The first two steps of the user start in the page P2.

The damping factor is subtracted from 1 and the result is divided by the number of pages (N) and this term is then added to the product of the damping factor and the sum score of the incoming pertinences.

$$f(x) = \frac{1-d}{N} + d \sum_N \frac{f(yi)}{L(yi)}$$

So the importance of any page comes largely from the relevance of other pages. The damping factor adjusts the derived value downward to limit the weight of votes.
Page X receives links transmitted by pages y1 ... yn, the parameter d is adjusted between 0 and 1.
(We usually give d the value0.85 .)
Similarly, L(y) is defined as the number of links issued by page y (outgoing links).

Diagram 1: A website structure

In this diagram, each page (P1 .... P11) is a web page. An arrow represents a link from one page to another one. The page size is proportional to the importance of the web page in the eyes of googlebot[6]. We note then that page 9, which has the most incoming links, has the largest size.

3. CALCULATION BY ITERATION

The importance of a page link A depends on the importance of the pages that emit a link to A, and therefore cannot be determined without knowing the importance of the latter, and all those that emit a link to them, and so on ... moreover, inter pages links can loop.
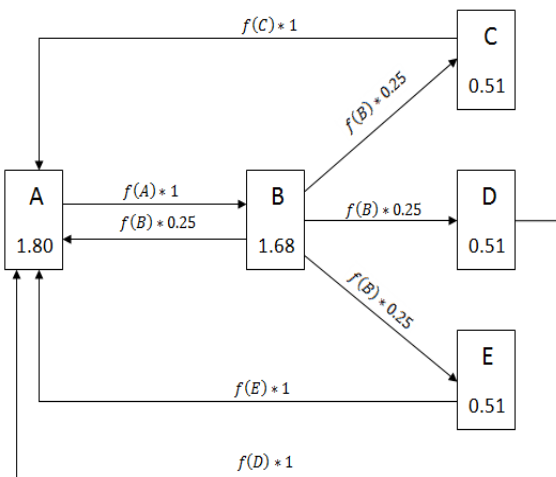
Diagram 2: A website loop

In this example, the home page (A) sends a link to the site map (B). The latter, in addition to returning the link to the home page, sends a link to each page of the site (C, D and E). To avoid "leakage" of Page Rank, they emit a link back to the homepage.

The sum of the PR of our pages being 5, we have then no loss due to incorrect chaining.

```
d = 0.85
N: number of incoming links
Yi: the source links pages
L (y): number of outgoing links
```

Example: f(c)= (1-0,85)/1+(0,85)*(1,68)/4 ➔0,51

The importance of a link can be calculated using a simple iterative algorithm.
This means that the calculation of the importance of a page can be done without knowing the importance of the final pages emitting a link to it.
This may seem paradoxical; nevertheless each iteration make results converge to a more and more precise value. The only thing to do is to retain the value obtained in order to start the next iteration with the latter.

Let's reinvent the Web in its simplest form:
2 pages A and B pages pointing towards one another.
Each page has an outgoing link, so L(A) = L(B) = 1
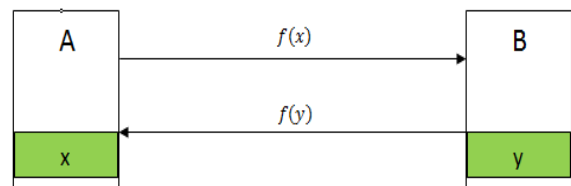
We do not know the relevance of the two pages, so we need a starting value, here the relevance of each page equals 0.

$$\begin{cases} f(A) = \frac{1-d}{N} + d \sum_N \frac{f(Bi)}{L(Bi)} \\ N=1 \quad , d= 0.85 \end{cases}$$

First iteration

$f(A) = 0.15 + 0.85 * 0 \quad = \mathbf{0.15}$
$f(B) = 0.15 + 0.85 * 0.15 = \mathbf{0.2775}$
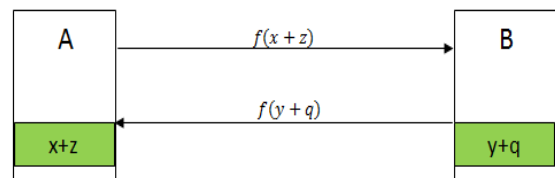
Second iteration

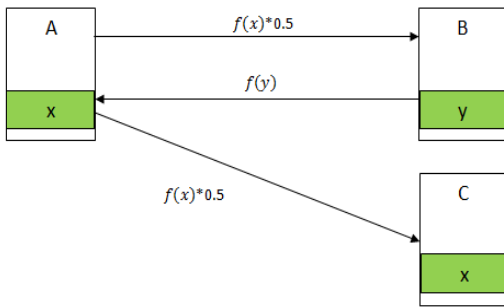$f(A) = 0.15 + 0.85 * 0.2775 \quad = \mathbf{0.385875}$
$f(B) = 0.15 + 0.85 * 0.385875 = \mathbf{0.47799375}$

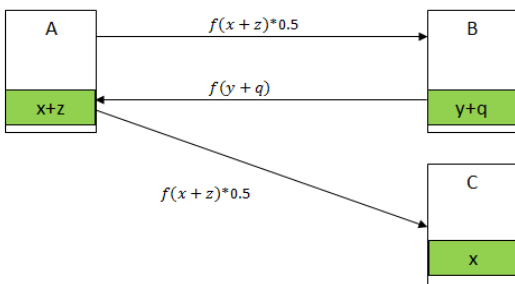We note that the values increase at each iteration.

Another example with 3 pages:

$$\begin{cases} f(A) = \frac{1-d}{N} + d \sum_N \frac{f(Bi)}{L(Bi)} \\ N=2 \quad , d= 0.85 \end{cases}$$

## First iteration

Second iteration

The mere fact of sending a link to an external page, from the beginning, makes the importance of the pages fall down significantly on all pages of the site.

The sole cause of the fall of the importance of the pages comes from the fact that page C returns no importance to Page A anymore.

So if a part of importance is redistributed to an internal or external page, it makes no difference if there is no back link to other pages
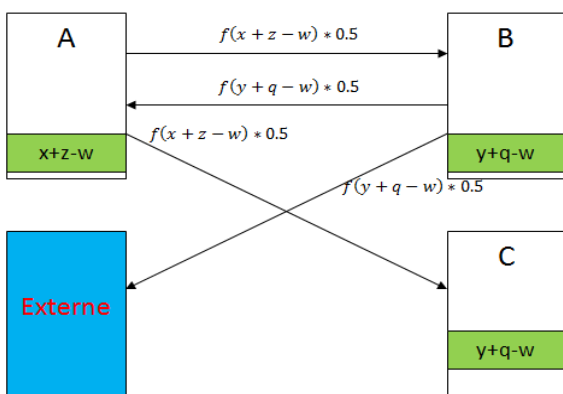

After calculating[7] (n iterations) the importance for each one of our sites, we found the following results:
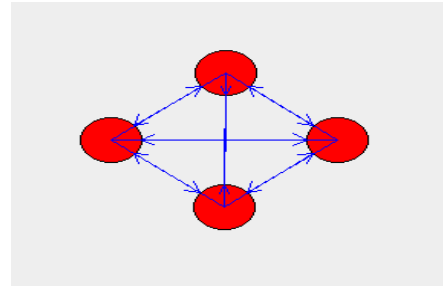
Site 1: according to the structure of the site, all pages are linked to each other, which mean that no page is predominant. We get the same relevance for all pages, while increasing ease of navigation for visitors. This type of chaining quickly becomes difficult to achieve as the number of pages on the site increases
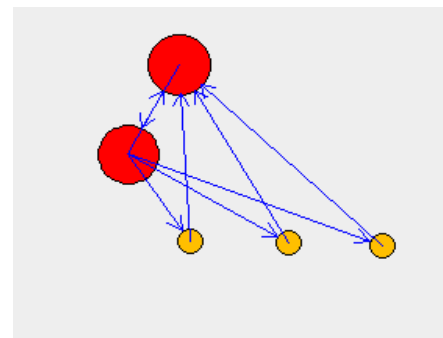
Site 2: the home page is the most important in this site; moreover we have no loss of relevance because of a bad chaining
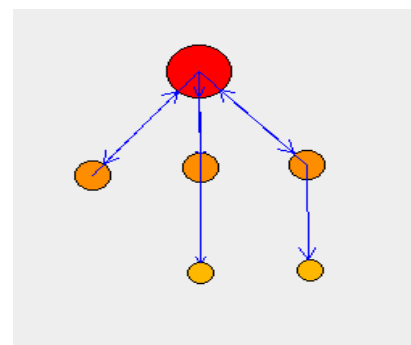
Site 3: The mere fact of sending a link to a page that returns no link drops the relevance consistently on all pages of the site

4. CONCLUSION

Based on the results obtained in the experiments, we conclude that the site structure 2 is the most feasible and performing for sites with a large number of pages; moreover, it presents a very effective method to get a high relevance in order to have a good SEO.

Links are very important for SEO, but it is not all that matters; in reality the content and manner in which the site is coded infects its position considerably during a search, so it is necessary to take a good time to consider the best way to work so as to achieve the site.

Also, be careful if you create links in large numbers, Google does not like it at all! This can have a negative impact on your

SEO. So create smarter web links if you want to reference a website. SEO is a long term work.

A good link should come from a page of the same content as that of the destination page. The opposite is true: make links to pages on the same theme.

It is advisable to include in the pages with duplicate content Canonical tag pointing to a page to keep. Here is the syntax:

<link rel="canonical" href="http://votresite/pagefinale"/>

5. REFERENCES

[1] D.R.W. Holton, I. Nafea, M. Younas, I. Awan . A class-based scheme for E-commerce web servers: Formal specification and performance evaluation. Journal of Network and Computer Applications, Volume 32, Issue 2, March 2009, Pages 455-460.

[2] Olivier Andrieu, Réussir son référencement web : Stratégie et techniques   SEO, Eyrolles - Edition 2014 (16 décembre 2013)

[3] Faisal Nabi. Secure business application logic for e-commerce systems   Original Research Article Computers & Security, Volume 24, Issue 3, May 2005, Pages 208-217

[4] Isabelle Canivet-Bourgaux, Référencement Mobile : Web analytics & stratégie de contenu, 456 pages Editeur : EYROLLES (11 juillet 2013)

[5] Samir Ghouti-Terki, Cookbook Référencement Google - 80 recettes de pros , (2 octobre 2013)

[6] Noel Nguessan, Bien référencer son site internet sur Google: L'Essentiel du référencement web; Noel Nguessan (8 septembre 2013)

[7] Yan Erjia, Topic-based Pagerank: toward a topic-level scientific evaluation, Scientometrics (Online) A. 2014, vol. 100, n° 2

[8] Springer, On the multi-splitting iteration method for computing PageRank, Journal of applied mathematics and computing (Internet) A. 2013, vol. 42, n° 1-2, pp. 479-490

[9] ERMANN Leonardo ; FRAHM Klaus M. ; SHEPELYANSKY Dima L; Spectral properties of Google matrix of Wikipedia and other networks; The European physical journal. B, Condensed matter physics (Print) A. 2013, vol. 86, n° 5

[10] Polyak B. T. ; Tremba A. A; Regularization-based solution of the PageRank problem for large matrices; Automation and remote control (Online) A. 2012, vol. 73, n° 11

[11] http://professeurs.esiea.fr/wassner/?2007/06/03/74-l-algorithme-pagerank-comment-a-marche

[12] http://en.wikipedia.org/wiki/PageRank