UTILIZAÇÃO DO ALGORITMO PAGERANK PARA OBTENÇÃO DE QUALIDADE NO RANQUEAMENTO DE PÁGINAS WEB

Cirene Assis Machado, Cristiane Mendanha Macedo Rocha, Eduardo José Magalhães, Maria das Dores Pereira e Silva

eduardo.magalhaes@ueg.br, cireneassis2009@hotmail.com, dodora@hotmail.com, cristianeK68@hotmail.com

Universidade Estadual de Goiás – Câmpus Goianésia – Sistemas de Informação

Goianésia – GO

RESUMO

Este artigo tem como objetivo apresentar uma métrica utilizada pelo Google dentro do seu algoritmo conhecida como *PageRank* (PR). A métrica foi criada a partir da necessidade de uma nova forma de busca, sendo utilizada para posicionar sites ou páginas entre os resultados, aprimorando a ideia do Altavista passando a levar em conta outros fatores, além da quantidade, evitando ao máximo a manipulação de resultados. O algoritmo baseia-se principalmente na contagem de *links* que apontam para determinada página na internet, aplicando a ponderação de pesos, evitando que mecanismos mal intencionados possam burlar o motor de busca. O *PageRank* determina o valor da página pelo número e pela qualidade dos *links* que apontam para cada página. Procurou-se determinar nesse estudo os fatores que influenciam na medição de qualidade de uma página. A pesquisa bibliográfica compreendeu estudos do algoritmo *PageRank,* através da aplicação de autovalores e autovetores e cadeias de Markov, prosseguindo com uma simulação utilizando sites para cálculo do PR, mostrando os resultados obtidos. Como resultado concluiu-se que a otimização de sites advém não somente do maior número de *backlinks* de sites externos, mas de outros fatores como popularidade, assim esses fatores podem contribuir para que os sites apareçam na primeira página do Google, o mais próximo possível dos primeiros lugares nos resultados de busca.

Palavras-Chave – Algoritmo, Google, Links, PageRank, Ranqueamento.

ALGORITHM PAGERANK USE FOR QUALITY IN OBTAINING RANKING WEB PAGES

ABSTRACT

This article aims to present a metric used by Google in its algorithm called PageRank (PR). The metric was created from the need for a new form of search and is used to locate sites or pages in the results, enhancing the idea of Altavista going to take into account factors other than the amount, so as to avoid the match-fixing . The algorithm is based primarily on the link count that link to a

particular page on the Internet, applying the weighting weights, preventing malicious mechanisms can bypass the search engine. The PageRank determines the page value by the number and quality of links pointing to each page. It was assessed in this study the factors that influence the quality of a page measurement. The bibliographic research included studies of the PageRank algorithm, by applying eigenvalues and eigenvectors and Markov chains, continuing a simulation using websites to calculate the PR, showing the results. As a result it was concluded that the site optimization comes not only the largest number of external sites backlinks, but other factors such as popularity, so these factors can contribute to the sites appear on the first page of Google, as close to the first places in the search results.

KEYWORDS – Algorithm, Google, Links, PageRank, Ranking.

I. INTRODUÇÃO

A partir da popularização dos computadores, e principalmente do advento da internet, o homem começou a gerar informações por meio digital de forma diversificada. A imensidão de páginas existentes hoje na web denota uma pequena parte deste grande repositório de informações quase ingerenciáveis. Nesse contexto, os algoritmos de ranqueamento têm sido utilizados para melhorar os resultados da busca capturando as informações relevantes, desempenhando um papel fundamental de busca, aprimorando a precisão dos resultados.

Para que as informações disponíveis na *web* se tornem acessíveis e úteis, é necessário a existência de um excelente serviço de busca, caso contrário, encontrar um site específico sem poderosos mecanismos, pode ser muito difícil ou até mesmo impossível.

Os primeiros mecanismos de busca na internet se baseavam em informações de suas páginas, e como o crescimento das páginas tende a ser exponencial, o mecanismo era insuficiente por conter páginas com mesmo conteúdo. Durante esse período foi desenvolvido o *PageRank*, onde a classificação das páginas baseava-se na centralidade do autovetor associado aos seus links.

Existem, entretanto, vários desafios para os mecanismos de busca, entre eles a enorme quantidade de informações a ser ranqueada e categorizada, a disponibilidade da base de dados, a qualidade nos resultados da busca, além da inexperiência dos usuários. O Google mantém uma lista de bilhões de páginas em ordem de importância, isto é, cada página tem sua importância na *web* como um todo, possuindo um banco de páginas que reúne desde as páginas mais relevantes e acessadas até as menos conhecidas. Essa importância dá-se pelo número de votos que uma página recebe. Um voto é um *link* em qualquer lugar da *web* para aquela página. Votos de páginas mais importantes tem um maior valor.

Neste artigo, será apresentado uma métrica utilizada nos algoritmos de ranqueamento do Google permitindo a classificação de páginas da Internet, a partir de apontamentos de outras páginas. Será feita uma abordagem da aplicação de autovalores e autovetores e cadeias de Markov, além de mostrar a contribuição das teorias que influenciam na popularização de um site diante nas buscas do Google. Será fator de avaliação a adoção dessa prática e a contribuição para que um site apareça na primeira página do Google, o mais próximo dos primeiros lugares nos resultados de busca.

II. FUNDAMENTAÇÃO TEÓRICA (1)

A. O Algoritmo *PageRank*

O PageRank é uma família de algoritmos de análise de rede que dá pesos numéricos a cada elemento de uma coleção de documentos hiperligados, como as páginas da Internet, com o propósito de medir a sua importância nesse grupo, por meio de um motor de busca. De acordo com Magalhães (s.d., p. 06) "O *PageRank* é apenas um dos métodos que o Google usa para determinar a relevância de uma página ou importância". O processo do algoritmo *PageRank* foi patenteado pela Universidade de *Stanford* nos Estados Unidos e apenas o nome é uma marca registrada da empresa *Google Inc*.

Desenvolvido por Larry Page e Sergey Brin, que são os fundadores do Google, o *PageRank* é um super conjunto de fórmulas matemáticas, que têm por objetivo mensurar a relevância de

determinadas páginas, para que elas possam ser posicionadas no motor de busca do próprio Google. Trata-se do medidor oficial da empresa. O *PageRank* serve para mostrar o conceito que determinada página tem dentro da sua área de atuação, baseando-se numa comparação feita entre a página e sites com conteúdos semelhantes. Assim define qual conteúdo será mais interessante para o usuário, de acordo com Rezende (2016).

B. Ranqueamento dos Conteúdos no Buscador

Na prática, o *PageRank* elege por meio de votos os conteúdos a serem posicionados no buscador. De acordo com Magalhães (s.d, p. 11), sendo um algoritmo de análise de *links*, atribui um peso numérico a cada elemento de um conjunto de *hiperlinks* de documentos, com o objetivo de "medir" a importância relativa dentro do conjunto. Sendo utilizado em diversas aplicações de qualquer coleção de entidades com citações e referências, tendo o peso numérico atribuído a qualquer elemento determinado, sendo também chamado de *PageRank* de E e é notado como PR(E).

Segundo Brin e Page (1998) quanto maior a importância da página, maior valor ela terá para o *PageRank*, tendo maior importância a partir de várias referências de outras páginas. No caso de páginas com maior *PageRank* possuírem *links* apontando para determinado site, maiores serão as chances dessa página subir no *ranking* do algoritmo. Assim, o simples fato de vários sites de menor relevância apontar para um determinado site não o torna importante.

Portanto, para aumentar o *PageRank* de determinado site, basicamente é necessário que mais *links* estejam apontando para ele, porém, os *links* apontados devem estar bem classificados para contribuir no ranqueamento. É importante também que os *links* desses sites tenham qualidade e sejam semanticamente relacionados ao conteúdo do *site*, assim poderão agregar valor ao *PageRank*. Para qualificar as páginas o mecanismo utiliza o sistema de pontuação numérica de 0 (zero) a 10 (dez). Normalmente uma pontuação em torno de 5 (cinco) já significa que o site está muito bem colocado. Segundo Júnior e Gallina (s.d, p. 02), "a inovação do *PageRank* vem de casos em que a contagem de citações não reflete a noção de importância que os usuários estão procurando". Nesse caso, um site de grande importância, como Yahoo, por exemplo, apontando uma única vez para um site desconhecido pode gerar a esse site uma posição maior no ranking, isso pelo peso do site apontador. Assim, o *PageRank* aproxima a importância da página aos *hiperlinks*.

C. O Processo de Distribuição de Valores

O algoritmo denominado *PageRank* gerou uma revolução nas tecnologias de recuperação de informação e no panorama de motores de busca conhecidos até o final dos anos 90. Pela primeira vez a imensidão de dados que circulavam na Internet passou a ser classificados e distribuídos pela Google em hierarquias dinâmicas, conforme a visibilidade e a importância de cada página. Para Braz e Katague (2013, p. 04) "Antes de 1998, não haviam mecanismos de busca que levavam em conta a estrutura de hyperlinks da rede, no entanto, alguns pesquisadores, como Larry Page e Sergey Brin já tinham ideias de como retirar informações desta estrutura da web."

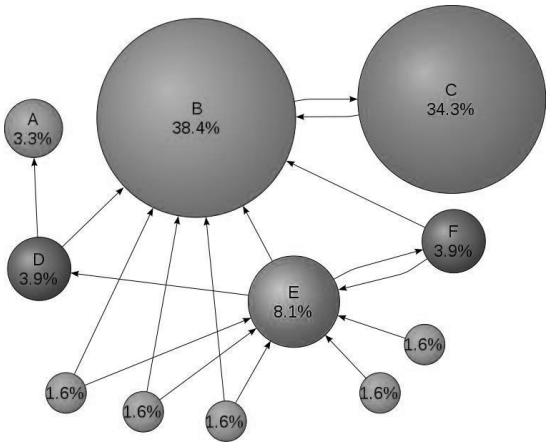
Esse ranqueamento das páginas web pode ser entendido de forma intuitiva como a determinação de valor para uma página a partir do número e pela qualidade dos *links* que procuram a página. Um *link* apontado, vindo de um endereço classificado no topo da lista das páginas

ranqueadas vale mais que um *link* apontado de uma página classificada na região inferior da lista das páginas ranqueadas. Segundo o Google (s.d) "O *PageRank* é a medida da importância de uma página com base nos *links* de entrada de outras páginas. Em outras palavras, cada *link* para uma página no seu site proveniente de outro site adiciona um *PageRank* ao seu *site*."

Enquanto, no final dos anos 90, motores de busca ainda classificavam a mão às páginas da web, organizando-as em estrutura de árvore, típica da estrutura do conhecimento enciclopédico, os criadores do Google já utilizavam métricas para localizar e atribuir um valor semântico a qualquer hipertexto, por mais dinâmico e caótico que fosse.

O *PageRank* começou descrevendo as páginas *web* segundo a sua popularidade, com o motor de busca devolvendo uma hierarquia de resultados, conforme o critério de ranqueamento usado. Segundo o SITE DO GOOGLE (s.d) "Os melhores tipos de links são aqueles retornados com base na qualidade do conteúdo.". As fórmulas matemáticas do algoritmo *PageRank* são consideradas altamente complexas, sendo entendidas apenas por matemáticos profissionais. Em seu funcionamento, o *PageRank* forma uma distribuição probabilística sobre todas as páginas da web de forma que a soma dos *PageRanks* de todas as páginas será sempre 1. A figura abaixo mostra como se dá essa distribuição de valores entre as páginas.

Fig. 1: Distribuição de valores entre as páginas



Fonte: (PASQUINELLI, 2009, p. 04)

Segundo Júnior e Gallina (s.d, p. 03) "O algoritmo do PageRank pode ser aplicado para uma coleção de documentos de qualquer tamanho", ainda para Júnior e Gallina (s.d) o valor do *PageRank* passado a partir de cada página seria distribuído entre os *links* da página, onde a página B passaria um PR de 0,125 para a página A e os outros 0,125 para a página C, e a página D passaria um terço de seu *PageRank* para a página A.

Onde:
PR(A) - PageRank da Página de Internet A
PR(B) - PageRank da Página de Internet B

PR(C) - PageRank da Página de Internet C
PR(D) - PageRank da Página de Internet D

Considerando o número de *links* de uma determinada página, retornado pela função L, o *PageRank* de A pode ser indicado como:

Onde:
L(2) - Número de links da Página de Internet 2
L(3) - Número de links da Página de Internet 3
L(4) - Número de links da Página de Internet 4

De um modo geral, podemos obter o *PageRank* de uma dada página u, sendo Bu o conjunto de páginas que apontam para u, dividindo a soma do valor do *PageRank* de cada página v pelo número de links que parte da página v:

Onde:
PR(u) - Dada página u
- Somatório do conjunto de páginas
PR(v) - PageRank de cada página
L(v) - Número de links de determinada página

Para tanto é usado um fator de amortização, que leva em conta a possibilidade de um *web surfer* aleatório, ou seja, um personagem qualquer clicar em *links* sucessivos aleatoriamente, pulando de uma página a outra. O fator de amortização *d* assume valores entre 0 e 1, sendo que o valor *d* é fixo em 0,85. A representação probabilística de se clicar *link* em uma página que está sendo visitada é de 85%, contra 15% de probabilidade de se escolher outra página aleatória para começar a navegar de novo.

Onde:
PR(A) - Representação probabilística para uma página
- fator de amortização

III. DESENVOLVIMENTO

O artigo desenvolvido trata de uma pesquisa bibliográfica realizada no período de Março de 2015 a Novembro de 2015, por meio de consulta a livros, revistas e artigos de sites da internet com relevância na área. Esse estudo foi realizado utilizando como método a pesquisa bibliográfica, de natureza qualitativa. A pesquisa teve como campo de ação, sobretudo a busca virtual em sites conceituados como Google e USP. Para a coleta de dados foram utilizados os seguintes descritores:

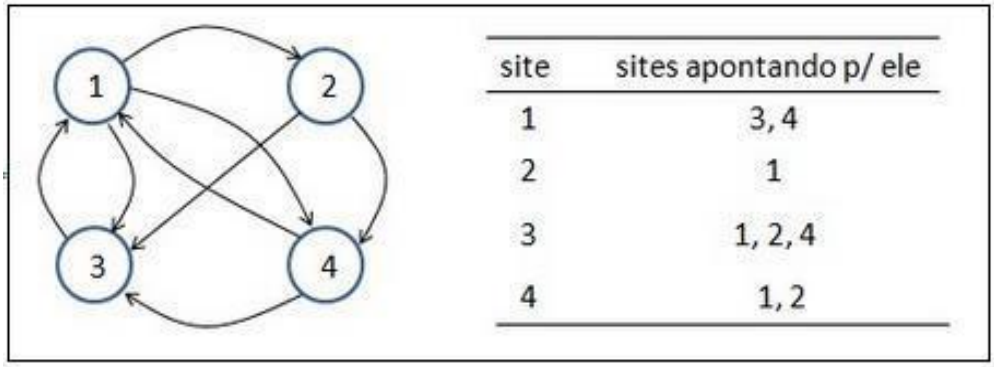
PageRank, Algoritmos de Classificação e Ranqueamento do Google.

A. Interpretação Probabilística

Devido à importância das informações, sobretudo nas últimas décadas com o armazenamento eletrônico e o crescimento das páginas web, o ranqueamento torna-se de extrema importância para a classificação, principalmente pelo valor da informação e a necessidade de encontrá-la de forma rápida, coesa e exata.

Para entender o funcionamento do algoritmo é necessário entender o processo da interpretação probabilística, tratado pela cadeia de Markov. A cadeia de Markov é um processo estocástico caracterizado por seu estado futuro depender apenas do seu estado atual, sendo que os estados passados não influenciam no estado futuro. De acordo com Filho (s.d) "A ideia fundamental do algoritmo é estabelecer um índice de popularidade para cada site da internet, fundamentado nos sites com *links* apontando para esse site, ponderado pelo índice de popularidade, e assim recursivamente". Observe a figura abaixo com a ideia fundamental do *PageRank*.

Fig. 2: Demonstração da ideia fundamental do PageRank

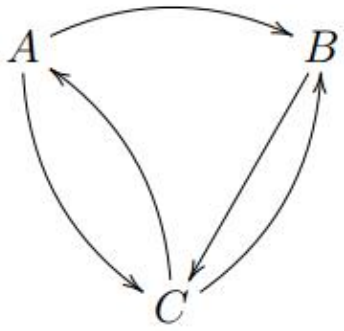


Fonte: (BRIAN et al., 2007 apud FILHO, s.d)

A partir da figura acima, conclui-se facilmente que o site C é o mais popular entre os quatro analisados, pois a *priori* tem-se que os sites (A, B e D) estão apontando para ele. O problema dessa avaliação é que foi considerado hipoteticamente apenas o número de *links* ou apontamentos de sites para avaliação de popularidade, ignorando a qualidade do conteúdo, até mesmo porque a simples criação de páginas, sem expressão alguma, poderia contribuir nesse caso para a alta popularidade de determinada página no ranqueamento do Google. Essa forma ingênua era usada nos primeiros algoritmos usados por buscadores, e que logo foram esquecidos.

De forma geral, a princípio consideramos a existência de apenas três sites, representados por A, B e C, conforme mostra a figura 2 abaixo

Fig. 3: Modelo básico de uma mini internet

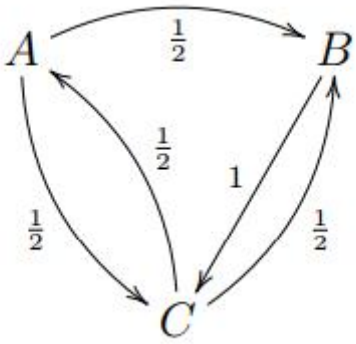


Fonte: (GOLMAKANI et. al, 2014, p. 21)

A partir do modelo proposto, é necessário levar em consideração que dentro de cada um desse sites haja *links* direcionados ao site respectivo. Torna-se fácil identificar qual dos sites seria o mais popular, ou melhor, qual destes deveria aparecer no topo de uma pesquisa em um buscador. A partir da contagem dos apontamentos, levando em consideração que B tem dois *links,* então, B teria dois votos, advindos de A e C, assim como C, também com dois apontamentos de A e B, ficando A com apenas um voto advindo de C.

Assim, pode-se dizer que B e C estariam empatados com dois votos em relação a sua popularidade. Contudo, na vida real esse método simples não pode ser aplicado por vários motivos. Uma solução inicial para o problema desenvolvido pelo Google foi, levar em consideração que os usuários estando no site A, tem a mesma probabilidade de acessar o site B ou C. Nesse caso observe a figura abaixo

Fig. 4: Solução inicial do Google



Fonte: (GOLMAKANI et. al, 2014, p. 22)

A partir das observações iniciais e montando uma matriz de transição de uma Cadeia de Markov regular, percebe-se que a probabilidade de que usuários acessem o site A depois de 16 cliques em links é igual a 0,222, que também é a mesma probabilidade para acesso a B ou C, após a mesma quantidade de cliques. Desta forma, a probabilidade para A depois de 16 cliques em links é 0,2, enquanto que para B e C são iguais a 0,333 e 0,444 respectivamente. Assim, tem-se que o site C é o mais popular e A é o menos popular.

IV. RESULTADOS

O *PageRank* é um algoritmo eficiente para ranqueamento de sites usado pelo Google, apesar de haver vários fatores que podem influenciar na popularidade de um site, em linhas gerais, quanto maior o número de sites que apontam para um site (ou seja, backlinks) maior será o seu *PageRank*. No caso do Twitter, por exemplo, existem muitos sites apontando pra ele, além de existir um número alto de *backlinks* de sites externos, com domínio .EDU e .GOV. O próprio *PageRank* para o site google.com no ano passado era inferior ao *PageRank* do twitter.com. Isso mostra a complexidade do algoritmo, e a quantidades de fatores que podem influenciar na popularidade de um site, levando em conta o seu *PageRank*. A figura abaixo mostra a classificação do site twitter.com.

Fig. 5 - Google PageRank do twitter.com

Fonte: (NIGAM, 2015)

No caso do Twitter, a explicação mais óbvia, seria de que os usuários sempre colocam o link do Twitter em sua página, blog, empresa ou outras redes sociais. Ao contrário, o Google normalmente não tem o seu domínio mencionado em outros sites, até por ser um padrão de pesquisa e um endereço amplamente conhecido.

Ainda como resultado foi feito alguns testes com páginas através do Website Value Calculator no endereço www.page-rank-calculator.com, entre esses usp.br, ueg.br, uol.com e terra.com, o site retorna algumas informações úteis como última atualização, idade, tráfego global de classificação, domínio, estimativa e renda diária, além é claro do *PageRank.* No caso do site uol.com o *PageRank* listado foi de 8/10, em seguida o site terra.com com *PageRank* de 7/10. O site da ueg.br não obteve classificação no *PageRank,* segundo o site ficando com 0/10. Enquanto o site usp.br obteve *PageRank* 8/10. A figura abaixo mostra a avaliação para o usp.br

Fig. 6 - PageRank para o endereço usp.br

Fonte: (WEBSITE VALUE CALCULATOR, 2016)

V. CONCLUSÃO

Este artigo propôs um estudo sobre o algoritmo *PageRank*. Foram apresentados tópicos que remetem à estrutura fundamental do algoritmo, além de autores com pesquisas publicadas na área, ademais foi dado ênfase ao *PageRank*, oportunidade em que foi detalhado o ranqueamento dos conteúdos no buscador e o processo de distribuição de valores. Para exemplificar foram abordados tópicos sobre o cálculo do *PageRank*, isto feito a partir de ferramentas disponíveis na internet.

Pôde-se constatar que o *PageRank* utiliza-se da cadeia de Markov, e que a classificação ou ranqueamento depende de uma serie de fatores, que inclusive inibem o modelo baseado apenas no número de *backlink* como atributo principal de classificação. Assim, para que o site possa ser melhor classificado nas pesquisas dos buscadores é necessário *backlink* de sites mais populares diante do Google, além da necessidade de semelhança no conteúdo.

O estudo mostrou-se promissor e trouxe informações que deram maior compreensão à forma utilizada pelo Google para classificar os sites existentes, assim como os modelos matemáticos e probabilísticos, podendo influir de forma significativa na classificação de sites e mediante consultas a determinado endereços eletrônicos. Seria recomendável que futuramente houvesse pesquisas em empresas de e-commerce do país para saber se há influência do algoritmo *PageRank* nos sites, e se a melhor pontuação é considerada de alguma forma como diferencial.

REFERÊNCIAS BIBLIOGRÁFICAS

BRAZ, Caio de Moraes.; KATAGUE, Gustavo Perez. **Classificação de Conteúdo Web: Estudo Comparativo e implementações.** São Paulo, fev. 2013. Disponível em: < https://linux.ime.usp.br/~katague/files/monografia.pdf >. Acesso em: 18 de ago. 2016.

BORGES, Fábio; WERNECK, Veronica S. F. de. **Classificando Páginas em Redes Desconexas**

com PageRank. Disponível em: < http://www.sbmac.org.br/eventos/cnmac/xxxi_cnmac/PDF/206.pdf> Acesso em: 18 de ago. 2016.

BRIN, Sergey; PAGE, Lawrence. **The Anatomy of a Large-Scale Hypertextual Web Search Engine.** 1998. Stanford. Disponível em: < http://infolab.stanford.edu/~backrub/google.html> Acesso em: 18 de ago. 2016.

FILHO, Adriano Azevedo. **Introdução ao Algoritmo PageRank do Google com o R: Uma Aplicação de Autovalores/Autovetores e Cadeias de Markov**. Disponível em: <https://rpubs.com/adriano/PageRank> . Acesso em: 10 de out. 2015.

GOLMAKANI et. al. **Cadeias de Markov**. 2014. Minicurso (Bienal da Sociedade Brasileira de Matemática) - Instituto de Matemática, Universidade Federal de Alagoas. 2014. Disponível em: < http://www.im.ufal.br/evento/bsbm/download/minicurso/cadeias.pdf >. Acesso em: 18 de agosto de 2016.

JÚNIOR, Roberto R.; GALLINA, Leandro Z. **PageRank para Ordenação de Resultados em Ferramenta de Busca na Web.** Disponível em: < http://www.inf.ufrgs.br/~lzgallina/files/Pagerank%20para%20Ordenacao%20de%20Resultados%20em%20Ferramenta%20de%20Busca%20na%20Web.pdf >. Acesso em: 31 de ago. 2015.

MAGALHÃES, Teresinha Moreira de. **O Motor de busca Google e o Algoritmo PageRank.** Disponível em: < http://fsd.edu.br/revistaeletronica/arquivos/6Edicao/artigo35%20TERESINHA.pdf >. Acesso em: 31 de ago. 2015.

NIGAM, Rohit. **Why is Google's Pagerank 9/10 wheras Twitter's Pagerank is 10/10?**. Ago. 2015. Disponível em: <https://www.quora.com/Why-is-Googles-Pagerank-9-10-wheras-Twitters-Pagerank-is-10-10>. Acesso em: 18 de ago. 2016.

PASQUINELLI, Matteo. **O algoritmo do PageRank do Google: Um diagrama do capitalismo cognitivo e da exploração da inteligência social geral.** Disponível em: <http://matteopasquinelli.com/docs/Pasquineli_PageRank_pt.pdf>. Acesso em: 30 de ago. de 2015.

PEREIRA JUNIOR, E. A. Google: Ferramenta de busca de Informação na Web. **Saber Digital: Revista Eletrônica do CESVA**, Valença, v. 1, n. 1, p. 18-32, 2008.

REZENDE, Isabela. **PageRank 2016: O que é e como funciona.** Jun. 2016. Disponível em: <http://blog.ingagedigital.com.br/pagerank-2016-o-que-e-como-funciona>. Acesso em: 18 de ago. 2016.

RICOTTA, Fábio. **O que é PageRank?** Mai. 2014. Disponível em: <http://www.agenciamestre.com/marketing-digital/o-que-e-pagerank/>. Acesso em: 18 de ago. 2016.

SITE DO GOOGLE. **Como a Pesquisa Google funciona**. disponível em: <https://support.google.com/webmasters/answer/70897?hl=pt-BR>. Acesso em: 18 de ago. 2016.

WEBSITE VALUE CALCULATOR. 2016. Disponível em: < http://www.page-rank-calculator.com/>. Acesso em 18 de ago. 2016.