

O MOTOR DE BUSCA GOOGLE E O ALGORITMO PAGERANK

Teresinha Moreira de Magalhães¹

RESUMO

O presente trabalho apresenta a metodologia de busca e classificação de páginas na Web utilizadas pelo motor de busca Google. Técnicas de recuperação das informações, bem como as características do algoritmo PageRank são apresentadas.

PALAVRAS CHAVES: Google, PageRank, algoritmo, WebCrawler

INTRODUÇÃO

Google PageRank - Todos usam, mas poucos sabem como ele funciona. Google *PageRank* é, provavelmente, um dos algoritmos mais importantes já desenvolvidos para a Web. Com bilhões de páginas existentes e milhões de páginas geradas a cada dia, a questão de pesquisa na Web é mais complexa do que, provavelmente, possa parecer. *PageRank* é apenas um dos centenas de fatores utilizados pelo Google para determinar os melhores resultados de busca, além de manter as buscas limpas e eficientes. Mas como é, realmente, feito? Como funciona o Google *PageRank*? Quais fatores impactam sobre ele e quais não? E o que realmente se sabe sobre *PageRank*?

A pesquisa se norteou através de dezenas de sugestões e fatos selecionados sobre o *PageRank*. Reuniram-se trabalhos acadêmicos relacionados ao tema - como propostas científicas para melhores resultados dos motores de busca. Foi possível encontrar referências a conhecimentos matemáticos do *PageRank*, bem como

¹ Doutora em Sistemas Computacionais, área multidisciplinar da Engenharia Civil. Professora do Curso de Sistemas para Internet das Faculdades Integradas Vianna Júnior, professora e diretora Acadêmica da Faculdade de Santos Dumont.
e-mail: tmagalha@yahoo.com.br.

ferramentas úteis para usar o *PageRank* para analisar a classificação dos projetos web.

O *PageRank*, por sua vez, é o algoritmo usado pelo motor de busca Google para ajudar a determinar a relevância ou importância de uma página, de acordo com o número de vezes que a página é referenciada por outros sites. Interpreta o link de uma página para a outra como um voto das mesmas. Além disso, analisa o valor da página que dá o voto. Os votos dados por páginas importantes pesam mais e ajudam a tornar outras páginas importantes (BRIN & PAGE, apud ZANIER, 2006, p. 39). Diante disso, sabe-se que trocar links entre sites pode ajudar na classificação, principalmente, se o site aliado estiver com boa qualificação no *PageRank*.

1. *Web Crawlers*

Segundo Menczer (2007), in (LIU, 2007), *Crawlers* são programas que automaticamente vasculham páginas *Web* para colher informações que podem ser analisadas e mineradas em um local *on-line* ou *off-line*. *Web Crawler* é um agente, um *bot* que vai, de página em página, analisando o código HTML, extraíndo informações e os *links* para continuar a sua tarefa. Existem dezenas de utilidades para *crawlers*, inclusive em ações maliciosas, porém o uso mais comum é na construção de sistemas de busca de páginas (*Google, Yahoo Search, Microsoft Bing*, dentre outros).

Os *crawlers* utilizados por esses *sites* têm como tarefa indexar todo conteúdo possível na *Internet*. Existem outros tipos de *crawlers* que se focam em garimpar informações sobre um assunto em específico ou se focam em baixar somente páginas que sigam um determinado padrão de URL. Esse robô interage diretamente com a *Web*. Possui como função descobrir novos documentos na *Internet* de forma a torná-los consultáveis. Os *Crawler*, automaticamente, visitam páginas *Web*, leem-nas, copiam-nas e seguem os *hiperlinks* nelas contidos (ETZIONI, 1999).

Constata-se, também, que o *Crawler*, além de capturar e transmitir muitos *sites* simultaneamente e de forma eficiente, tenta prever a similaridade entre o conteúdo do arquivo e a consulta do usuário (MAGALHÃES, 2008).

Segundo Markov & Larose (2007), navegar na *Web* é um modo muito útil para explorar uma coleção de documentos *linkados* quando se conhece um tema ou área pela qual haja interesse. Entretanto, um *browser*, por si só, é incapaz de obter informações sobre determinado assunto ou tema. A melhor abordagem é ter páginas *Web* organizadas por tópico ou pesquisar uma coleção de páginas indexadas por palavra-chave.

Para tanto, é preciso levar em conta que rastreamento da *Web* envolve interação com centenas de milhares de servidores *Web*, concebidos não só para satisfazer diferentes objetivos, mas também para prover diversos serviços, tais como acessos ao banco de dados, interações dos usuários, geração de páginas dinâmicas, e assim por diante. Outro fator importante é o imensurável número de páginas que devem ser visitadas, analisadas e armazenadas. Assim, um *Web crawler* concebido para rastrear toda a *Web* é um sistema sofisticado que utiliza tecnologia avançada de programação. Portanto, para melhorar seu tempo e sua eficiência de espaço, é geralmente executado em computadores com processamento paralelo e de alto desempenho.

A seguir, um breve resumo dos problemas comuns enfrentados, em grande escala, pelos *crawlers*, bem como as soluções apresentadas, segundo Markov (2007). Deve-se ressaltar que, para não se afastar do objetivo principal, que é analisar o conteúdo da *Web*, o presente estudo não pretende detalhar pormenores técnicos.

O processo de buscar uma página da *Web* envolve uma rede latência (por vezes um "tempo"). Para evitar a espera do carregamento de uma página atual, para continuar com a próxima página, rastreadores buscam várias páginas simultaneamente. Por sua vez, essa busca exige a conexão com vários servidores (normalmente milhares), ao mesmo tempo, o que se consegue utilizando tecnologia de programação paralela e distribuída tais como *multithreading* (executando vários clientes, concomitantemente).

O primeiro passo para buscar uma página da *Web* é a resolução do endereço, convertendo o endereço simbólico da *Web* em um endereço IP. Isto é feito por um servidor DNS conectado pelo rastreador. Como várias páginas podem ser

localizadas em um único servidor, o armazenamento de endereços em um *cache* local permite que o rastreador evite a repetição de pedidos DNS em páginas já visitadas, conseqüentemente, melhorando a sua eficiência e minimizando o tráfego na *Internet*.

Após buscar uma página da *Web*, ela é digitalizada, e os URLs são extraídos. Transformam-se em *links* que serão seguidos pelo próximo *crawler*. Há muitas maneiras de especificar uma URL em HTML. Também pode ser especificado usando o endereço IP do servidor. Como o mapeamento entre o nome do servidor e os endereços IPs são muitos para muitos², isso pode resultar em vários URLs para uma única página da *Web*. O problema é agravado pelo fato de que *browsers* são tolerantes a páginas que contenham erro de sintaxe. Como resultado, são documentos HTML não escritos com cuidado e incluem, muitas vezes, erroneamente, URLs maliciosos, bem como outras estruturas. Tudo isso mostra que extrair URLs a partir de documentos HTML não é uma tarefa fácil. A solução é usar um bem concebido e robusto indexador e, após extração de URLs, convertê-los para a forma canônica. Mesmo assim, há armadilhas em que o rastreador pode cair. A melhor política é a de recolher estatísticas regularmente sobre cada um, rastrear e utilizá-los em um módulo especial chamado “um guarda”. A finalidade do guarda é excluir os *outlinks* provenientes de *sites* que dominam a coleção de *crawlers* das páginas.

Uma parte importante do sistema *Web crawler* é o repositório de texto. *Yahoo* argumentou que, em agosto 2005, seu índice incluiu 20 mil milhões de páginas. Com uma média de 10 KB para um documento da *Web*, este faz cerca de 200.000 GB (*gigabytes*) de armazenamento. Gerir esse incomensurável repositório é uma tarefa desafiadora. Note-se que este é o indexador repositório, e não o indexador de coleção de páginas da *Web* utilizadas para responder a consultas de pesquisa. Este último é de dimensão comparável, embora seja ainda mais complicado devido à necessidade de rápido acesso.

² Um servidor pode ter mais de um endereço IP, e diferentes nomes de *hosts* podem ser mapeados em um único endereço IP.

O repositório é utilizado para armazenar páginas rastreadas, manter em *cache* a URL e documentos necessários pelo rastreador, além de proporcionar o acesso para a construção de índices na fase seguinte. Para minimizar necessidades de armazenamento, as páginas *Web* são geralmente comprimidas, reduzindo os requisitos de armazenamento. Para *crawles* de grandes escalas, o repositório de texto pode ser distribuído para um número de servidores de armazenamento.

O objetivo de um *Web crawler* utilizado por um motor de pesquisa é proporcionar o acesso a páginas *Web* localmente. Isso significa que a *Web* deve ser rastreada regularmente para *update* das páginas. Tendo em vista a enorme capacidade do repositório de texto. A necessidade de atualizações regulares representa outro desafio para o *Web crawler*. O problema é o alto custo de atualização de índices. Uma solução comum é anexar as novas versões de páginas *Web* sem apagar as anteriores. Isso não só aumenta os requisitos de armazenamento, mas também permite que o repositório possa ser utilizado para efeitos de arquivo. Na verdade, existem indexadores que são utilizados apenas para os fins de arquivamento da *Web*. O mais popular da *Web* é o arquivo na *Internet Archive* <http://www.archive.org/>.

O Rastreamento da *Web* também envolve a interação de desenvolvedores de página. Como Brin e Page mencionaram em um *paper* sobre o seu motor de busca, o Google, eles estavam recebendo *e-mails* de pessoas comunicando visitas às suas páginas. Para facilitar essa interação, há normas que permitem servidores *Web* e *crawlers* trocar informações. Um deles é o robô *exclusion protocol*. Um arquivo chamado *robots.txt* que lista todos os caminhos prefixos de páginas que o *crawler* não deveria buscar é colocado no diretório raiz, HTTP do servidor, e lido pelos *crawles* antes do servidor principal.

Discute-se rastreamento em links e visitas de páginas sem levar em conta a sua semântica. No entanto, não se deve olvidar que a discussão no contexto da pesquisa é o rastreamento na *Web*. Assim, para melhorar a sua eficiência, ou para fins específicos, o rastreamento pode ser feito também como um *guided (informed) search*.

Geralmente, o *crawler* precede a fase de avaliação e classificação da página *Web*, por último vêm a indexação e a recuperação de documentos. No entanto, páginas *Web* podem ser avaliadas enquanto estão sendo rastreadas. Assim, há algum tipo de *crawler* que usa o método *ranking* para alcançar partes interessantes da *Web* e evita buscar páginas irrelevantes ou desinteressantes.

2. O que é o *PageRank*?

"O *PageRank* é apenas um dos métodos que o Google usa para determinar a relevância de uma página ou importância". "O Google usa vários fatores no ranking. Desses, o algoritmo *PageRank* pode ser o mais conhecido. Ele avalia dois aspectos: quantos links existem para uma página web a partir de outras páginas, e a qualidade dos sites vinculados. Com o *PageRank*, cinco ou seis ligações de alta qualidade, a partir de sites como www.cnn.com e www.nytimes.com, seriam muito mais valorizados do que muitos links de sites menos respeitáveis ou estabelecidos.

PageRank só foi uma aproximação da qualidade de uma página web e nunca teve nada a ver com a medida da relevância tópica de uma página web. Relevância tópica é medida com o contexto dos links e os fatores na página, como a densidade de palavras-chave, tags, títulos etc.

"*PageRank* usa a estrutura de links como um indicador do valor de uma página individual. Google interpreta um link da página A para a página B como um voto da página A para a página B. Google olha, consideravelmente, mais do que o volume de votos, ou links que uma página recebe. Por exemplo, analisa, também, a página que lança o voto. Os votos dados pelas páginas que são "importantes" pesam mais e ajudam a tornar outras páginas "importantes".

Cada link de entrada é importante para o total geral, exceto locais proibidos. O *PageRank* é uma forma de um sistema de votação. Um link para uma página é um voto para essa página. Páginas com maior *PageRank* são vistas pelo Google como mais importante. Seus votos são dados mais valor pelo Google - mas muito valor, em alguns casos. Em geral, quanto mais links de voto, mais forte é o *PageRank*.

A adição de novas páginas pode diminuir o *PageRank*. "O efeito é que, enquanto a *PageRank* total do site é aumentado, uma ou mais páginas existentes

sofrerão uma perda de *PageRank*, devido à nova página fazendo ganhos”. Até certo ponto, quanto mais páginas novas são adicionados, maior é a perda para as páginas existentes. Com grandes sites, esse efeito é pouco notado, mas, com as menores, provavelmente, o seria.

No Follow Treatment Search Engine Comparison

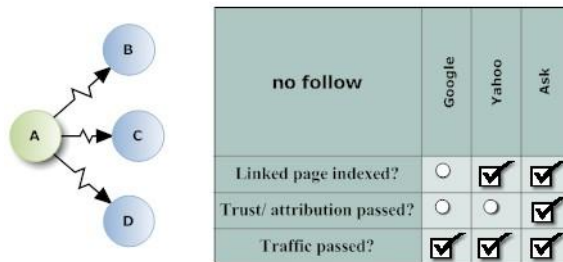


Figura 2. Google PageRank: Teoria Histórica e Científica

Para obter resultados de busca mais precisa, propõe-se um conjunto de vetores tendenciosos usando um conjunto de temas representativos, para captar com mais precisão a noção de importância no que diz respeito a um tema específico. Ao usar esses (pré-computados) vetores tendenciosos do *PageRank* para gerar escores de consultas específicas para as páginas no momento da consulta, é possível gerar classificações mais precisas do que com um único vetor de *PageRank*, genérico.

O método de classificação de links em um banco de dados atribui importância para nós em um banco de dados vinculados, como qualquer banco de dados de documentos que contenham citações, a *world wide web* ou qualquer outro banco de dados de hipermídia. A classificação atribuída a um documento é calculado a partir dos documentos citados. Além disso, a classificação de um documento é calculado a partir de uma constante que representa a probabilidade de um browser através do banco de dados contendo os documentos.

3. Como funciona o *PageRank*?

PageRank é apenas um dos inúmeros métodos usados pelo Google para determinar importância ou relevância de uma página. O Google interpreta os links de uma página para outra como um voto a página referenciada. O Google não analisa apenas o volume de votos, entre outros aspectos analisa também a página que favorece o voto. No entanto, esses aspectos não contam, quando o *PageRank* é calculado.

PageRank é baseado em links recebidos, mas não apenas sobre o número deles - pertinência e qualidade são importantes (em termos do *PageRank* dos sites que apontam para outro determinado site).

A equação que calcula o *PageRank* de uma página é determinada por:

$$PR(A) = (1-d) + d (PR(t1) / C(t1) + \dots + PR(tn) / C(tn)).$$

Nem todos os links tem peso iguais quando se trata de PR. Supondo uma página web com um PR8 e tenha um link sobre ela, o site linkado obterá uma quantidade justa do valor PR. Mas, se houver 100 links na página, cada link individualmente só obterá uma fração do valor. Vale lembrar ainda que links ruins recebidos não tem impacto sobre *PageRank*. O conteúdo não é levado em conta quando o *PageRank* é calculado.

PageRank não classifica sites da Web como um todo, mas cada página individualmente. Cada link de entrada é importante para o total geral. Exceto sites proibidos, que não contam. Os valores do *PageRank* não variam de 0 a 10. *PageRank* é um número de ponto flutuante.

A dificuldade de alcançar o *PageRank* é progressiva de acordo com cada nível. *PageRank* é calculado, numa escala logarítmica.

O Google calcula o PRs das páginas permanentemente, mas a atualização pode ser vista dentro de um período de tempo (Google Toolbar).

O sistema *PageRank* é usado pelo motor de busca Google para ajudar a determinar a relevância ou importância de uma página. Foi desenvolvido pelos

fundadores do Google, Larry Page e Sergey Brin enquanto cursavam a Universidade de Stanford em 1998.

O Google mantém uma lista de bilhões de páginas em ordem de importância, isto é, cada página tem sua importância na Internet como um todo; esse banco de páginas mantém desde a página mais importante até a menos importante. Essa importância se dá pelo número de votos que uma página recebe. Um voto é um link de qualquer lugar da Internet para aquela página. Votos de páginas mais importantes valem mais do que votos de páginas menos importantes.

Esse critério de ordenação das páginas, de acordo com várias pessoas, é bastante democrático, refletindo no que a "Internet pensa" sobre determinado termo. Lembre-se que cerca de bilhões de páginas são levadas em conta. A qualidade das páginas mais importantes são naturalmente garantidas, classificadas e eleitas pela própria Internet. Todas as páginas tem as mesmas condições de subir nessa lista, conquistando votos pela Internet.

Uma boa unidade de medida para definir o *PageRank* de uma página pode ser a porcentagem (%) que ela é mais importante. Por exemplo, se uma página tem *PageRank*TM de 33% significa que ela é mais importante que um terço de toda a Internet. Se o seu *PageRank*TM é 99% significa que ela é superior a quase todas as páginas da Internet.

No entanto, é possível manipular o *PageRank*TM atribuindo *links* descontextualizados com o objetivo da página, modificando a ordenação de resultados na pesquisa pelo Google e induzindo a resultados pouco relevantes ou tendenciosos. Um exemplo recente disso é a pesquisa por *failure* ou *miserable failure* que retornava como primeiro site a biografia oficial da Casa Branca para o presidente dos Estados Unidos, George W. Bush e em sequência a página de Michael Moore, inimigo declarado do presidente dos EUA. Este processo ficou conhecido por *Googlebombing*. Apesar do Google vir removendo alguns resultados decorrentes de "*Googlebombing*", este fato ainda acontece. A seguir uma figura que elustra as classificações das páginas pelo Googlebot.

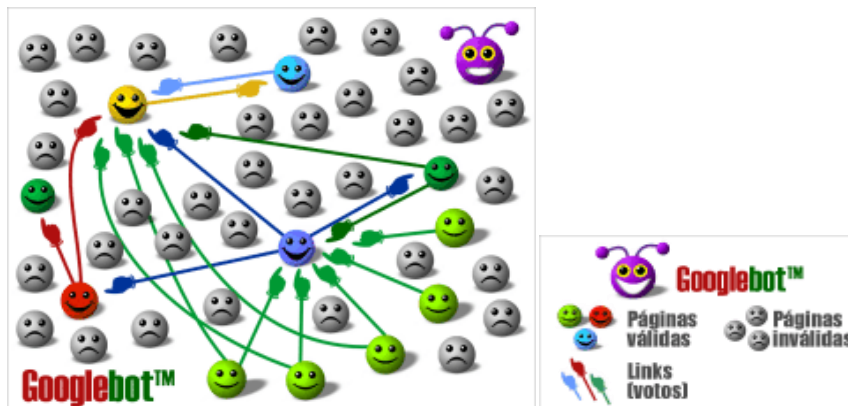


Figura 3: GoogleBot que varre as páginas calculando o PageRank™. Fonte: Wikipédia

Para verificar o *PageRank*™ de uma determinada página existem duas opções: Instalar a Google Toolbar que visita cada página e apresenta, imediatamente, o *PageRank*™ do site na própria barra ou visitar sites que fornecem a cotação do site digitado.

4. Impacto do *PageRank* sobre o Google

O alto *PageRank* não significa ranking elevado de buscas. Resultados DMOZ e Yahoo como sites .edu e .gov não melhoram o *PageRank* automaticamente.

Sub-diretórios não têm, necessariamente, menor *PageRank* que diretórios raízes e links marcados com atributo "*nofollow*" não contribuem para o Google *PageRank*.

Links eficientes em um site tem impacto sobre o *PageRank*. Sites relacionados a classificações elevadas tem mais peso. Mas, "uma página com *PageRank* elevado pode contar menos, se tiver muitos links demasiadamente espalhados." *Links* de e para sites relacionados de alta qualidade têm um impacto sobre *PageRank* e votos múltiplos para um *link* a partir de uma mesma página custa tanto quanto um único voto.

CONCLUSÃO

PageRank é um algoritmo de análise de *links*, o qual atribui um peso numérico a cada elemento de um conjunto de *hiperlinks* de documentos, tais como a World Wide Web, com o objetivo de "medir" a sua importância relativa dentro do conjunto.

O algoritmo pode ser aplicado a qualquer coleção de entidades com citações e referências recíprocas. O peso numérico atribuído a qualquer determinado elemento E é também chamado de *PageRank* de E e é notado como PR(E).

BIBLIOGRAFIA

BRIN, S. & PAGE, L., *The anatomy of a large scale hypertextual Web search engine*. *Computer Networks and ISDN Systems*, 1998.

ETZIONE, O. "*The World Wide Web Quagmire or gold mine*" *Communications of the ACM*, 1996.

FLORESCU, D.; LEVY, A.; MENDELZON, A. *Database Techniques for the World-Wide Web: A Survey*. ACM, New York, NY, USA, p. 59-74, 1998.

MAGALHÃES, L. H. **Uma análise de ferramentas para mineração de conteúdo de páginas Web**. Dissertação de M. Sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil, 2008.

MARKOV, Z.; LAROSE, D. T. *Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage*. New Britain: Wiley, 2007.

ZANIER, A. M. A. **A Evolução dos mecanismos de busca on-line: A melhoria nos resultados obtidos**. Dissertação de M. Sc., Faculdade de Economia e Finanças IBMC. Porto Alegre. Rio de Janeiro, 2006.