PAGE RANK: O FUNCIONAMENTO DA FERRAMENTA DE BUSCA DO GOOGLE

Marcos Vinicius Oliveira[1]
Joao Otavio Bovoloni[2]
Efraim Santana Leite Filho[3]
Gabriel Menezes[4]

Ciências da Computação

cadernos de graduação
ciências exatas e tecnológicas

ISSN IMPRESSO 1980-1777
ISSN ELETRÔNICO 2316-3135

RESUMO

Neste artigo, é apresentada a maneira que o Google inovou o seu motor de busca para responder os milhões de consultas por dia. Esta ferramenta auxilia seus usuários na obtenção de informação de forma rápida e concisa, evitando a pesquisa exaustiva, como nos textos físicos. Este trabalho científico propõe explicar os conceitos básicos e o funcionamento do algoritmo utilizado pelo *Google*, o *Page Rank*. Para fundamentação teórica, são explicados importantes conceitos da Álgebra Linear, como a Cadeia de Markov, método matemático utilizado pelo algoritmo para fornecer a informação de forma precisa. É explanado o modo que o algoritmo aproveita a lógica das probabilidades markovianas para ordenar os links, ao considerar a relevância da página e a importância da informação nela contida. Este trabalho científico procura despertar a curiosidade, por meio do conhecimento, ao esclarecer de forma didática e sucinta, sobre o mecanismo de pesquisa, já citado, mais popular do planeta.

PALAVRAS-CHAVE

Google. Page Rank. Álgebra Linear. Cadeia de Markov.

ABSTRACT

In this article, it's presented how Google innovated its search engine to handle millions of researches per day. This tool helps its users to get information in a fast and concise way, avoiding the exhaustive searches, like it is done in the physical texts. This paper proposes to explain the basics concepts and how the algorithm used by Google works, the Page Rank. For a theoretical validation, it's explained important concepts of Linear Algebra like Markov's chain, a mathematical method used by the algorithm to give to the user the required information. It's explained how the algorithm uses the Markov's properties logic to order the links, considering the page's importance and relevance of the information contained in it. This paper awakes the curiosity, by using knowledge, showing in a simple and didactic way, about the search engine, already cited, most popular in the planet.

KEYWORDS

Google. Page Rank. Linear Algebra. Markov's Chain.

1 INTRODUÇÃO

A *World Wide Web* tornou-se o principal marco da era da informação e "definiu" a morte da era industrial. Entretanto, mesmo com a revolução do armazenamento e acessibilidade de dados pela *web*, os usuários iniciantes ainda tinham dificuldade em realizar pesquisas por meio da internet. Isso mudou em 1998, quando os famosos mecanismos de busca começaram a analisar *links*, técnica essa que aperfeiçoaria a qualidade das informações requisitadas pelo usuário (LANGVILLE; MEYER, 2011).

A ferramenta de pesquisa do *Google* está presente na vida das pessoas que usam a grande rede de computadores. Ela auxilia os internautas na busca, ao facilitar, de forma eficiente e rápida, a obtenção de determinada informação por meio de um tipo de filtragem próprio e preciso. Isso é feito por meio de vários algoritmos dos quais um deles é chamado de *Page Rank*. Por ele o *Google* consegue selecionar *links* mais relevantes, que contêm as melhores informações para mostrar ao usuário.

O funcionamento do *Page Rank* se baseia na aplicação de parte da Álgebra Linear, utilizando principalmente o conceito de determinantes e da cadeia de Markov. Dessa forma, é possível identificar quais os sites são mais seletivos e posicionar, em ordem de relevância, os *links* com informações mais importantes para o assunto pesquisado. Assim, o tempo de pesquisa é altamente otimizado.

Este artigo tem como objetivo ser uma revisão de outros trabalhos, previamente publicados, sobre o assunto do algoritmo *Page Rank*, com intuito de explicar seu funcionamento e seus conceitos básicos de forma mais simplória. A fim de alcançar esses objetivos, serão utilizados como referencial teórico os processos estocásticos, processo markoviano e a cadeia de Markov.

2 PROCESSOS ESTOCÁSTICOS

Um processo estocástico descreve procedimentos de um sistema em um determinado período de tempo. Estabelecendo *X(t)* como uma função que muda seus valores ao decorrer do tempo, um processo estocástico tem valores aleatórios de acordo com o período passado como parâmetro. Os valores que *X(t)* pode assumir chamam-se estados e o seu conjunto *X*, espaço de estados (ALVES, 1997).

Os processos estocásticos podem ser classificados em relação ao estado, podendo ser contínuos, em sequência, ou discretos, em cadeia. Neste, é possível definir um conjunto enumerável ou finito, naquele, o caso é contrário. Em relação ao tempo pode ser classificado em discreto e contínuo, no primeiro o tempo é finito ou contável, já no segundo, é infinito ou incontável.

Há vários exemplos de processos estocásticos no cotidiano de quase todas as pessoas, a exemplo de *X(t)* representar um estado de uma máquina, podendo estar ligada ou desligada, no fim do dia *t* ou de *X(t)* representar o número de clientes em um restaurante no momento *t* (ALVES, 1997).

Como citado nos exemplos acima, é possível observar que o tempo pode ser percebido de forma discreta, no exemplo da máquina, ou de forma contínua, no do restaurante. Por definição, a variável *tempo* é contínua, mas esse conceito pode ser desfeito, se determinado acontecimento for observado em dado algum período.

2.1 PROCESSO MARKOVIANO

De acordo com Rui Alves e Catarina Delgado, em *Processos* Estocásticos (1997), um processo estocástico diz-se Markoviano, se for estacionário e gozar da propriedade de Markov ou da "perda de memória", isto é, apenas se o seu comportamento futuro depender do estado presente, independentemente dos estados visitados no passado. De fato, para um processo de Markov, é completamente irrelevante qualquer informação sobre estados passados, ou sobre o tempo de permanência no estado presente (ALVES, 1997)

2.2 CADEIA DE MARKOV EM TEMPO DISCRETO

Uma cadeia de Markov em tempo discreto é um processo estocástico em que a variável *t* representa intervalo de tempo contável ou finito, sendo que *{X(t)| t=0,1,2,3....}*, e que obedece a propriedade markoviana, ou seja, a probabilidade de *X(t)* estar no estado *j* no próximo momento depende apenas do estado atual, não dos estados passados. Fato que pode ser verificado na fórmula abaixo (ALVES, 1997)

$$P_{ij} = P\{X(t+1) = j \mid X(t) = i,\ X(t-1) = k_{t-1}, \ldots,\ X(1) = k_1,\ X(0) = k_0\} =$$
$$P\{X(t+1) = j \mid X(t) = i\},\ \forall\ t = 0, 1, 2, \ldots\ \forall i, j, k_0, k_1, \ldots, k_{t-1} \in X$$

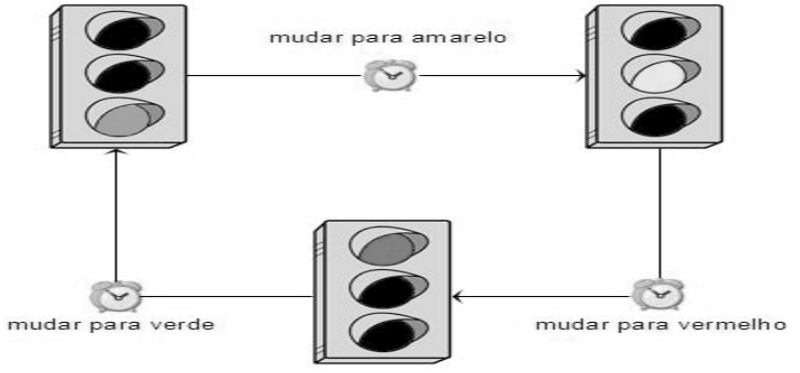
Uma cadeia de Markov pode ser representada por meio de um "diagrama de transições", como mostrado na Figura 1, ou por meio de uma matriz de transição *P* contendo as probabilidades de transição entre estados em determinado momento, como mostrado na *Matriz de transição* (Exemplo) (ALVES, 1997).

Na matriz de transição *P,* em que *i* representa estado atual e *j,* o seu estado futuro, a soma de cada linha deve ser igual á "1", ou 100%. O valor do elemento *Pij* representa a probabilidade de transição entre o estado i para o j em determinado período. Em um momento futuro, chegar-se-á a um determinado tempo em que haverá um estado de equilíbrio, ou seja, alcançar-se-á um ponto em que as mudanças serão insignificantes e as probabilidades irão permanecer as mesmas. Dá-se a isso o nome de *Ponto* de Equilíbrio ou *Estado Estacionário*.

Matriz de transição (Exemplo)

$$P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1j} & \cdots \\ p_{21} & p_{22} & \cdots & p_{2j} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ p_{i1} & p_{i2} & \cdots & p_{ij} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

Figura 1 – Diagrama de Transições

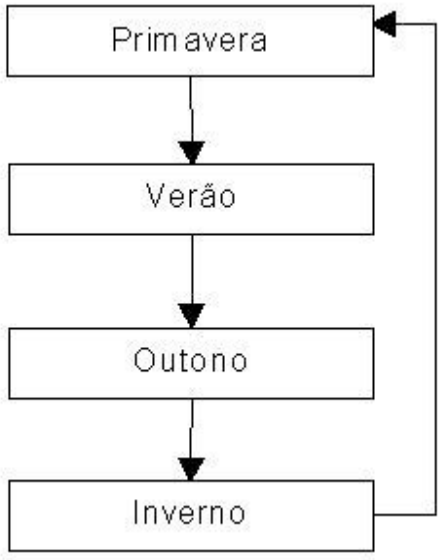


Fonte: Wikipedia (2016).

2.3 CADEIA DE MARKOV ERGÓDICA

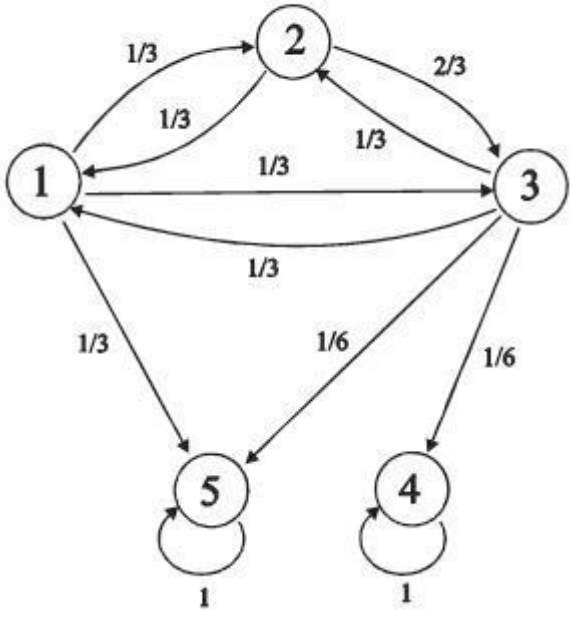
Uma cadeia de Markov, ao ser classificada ergódica, ou irredutível, caracteriza-se pelo fato de que qualquer estado pode retornar a qualquer outro em um número *n* definido de passos. Ou seja, esse tipo de cadeia não possui sumidouro, que é um estado onde não é possível alcançar qualquer outro.

Figura 2 – Cadeia ergódica ou irredutível



Fonte: Wikipedia (2016).

Figura 3 – Cadeia não ergódica ou redutível



Fonte: Amazonaws (2016).

Uma Cadeia de Markov não ergódica ou redutível, como mostrada na Figura 3, possui uma propriedade contrária à ergódica, ou seja, nela existem estados que não permite retornar a qualquer outro em um número *n* definido de passos. Assim, esse tipo de cadeia possui sumidouro.

3 PAGE RANK

3.1 MOTOR DE BUSCA

O *Google* foi criado por Larry Page e Sergey Brin, em 1997, na época do crescimento da *web*. No início, o algoritmo primeiramente escolhido foi o *BackRub*, que era hospedado na universidade de Stanford; contudo, em 1998, foi implementado o *Page* Rank, um novo método de pesquisa cujo nome homenageia um dos fundadores, código que prometia ser mais eficiente. Agora, não mais é necessário simular os livros na internet, os resultados desejados são quase que imediatamente apresentados, portanto, evita-se a pesquisa exaustiva, como nos textos físicos.

Figura 4 – Larry Page e Sergey Brin

Fonte: Scheidies (2016).

3.2 O ALGORITMO

A fim de fornecer ao usuário a informação requisitada, o *Page Rank* calcula a importância do site, analisando a qualidade dos links que apontam para outra página que, por sua vez, também será analisada, conhecidos como *back links*. Para um melhor entendimento desse processo, pode-se pensar que se uma pessoa de pouca relevância na área de tecnologia afirmar que a *Universidade Tiradentes* possui bons cursos dessa área, a opinião dela não tem muita importância; no entanto, se Bill Gates, dono da *Microsoft*, afirmar que esta instituição é uma das melhores, sua opinião terá

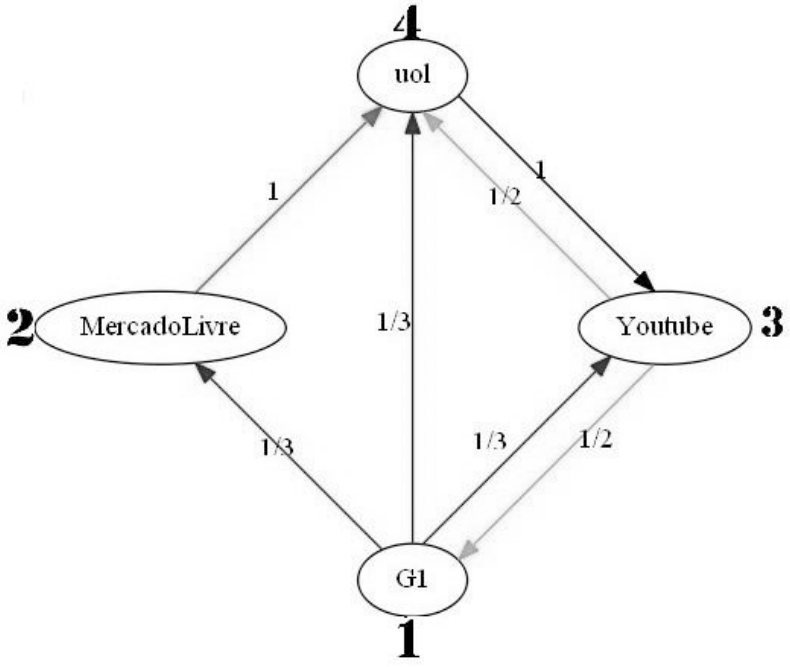
um impacto maior, visto que ele é uma pessoa importante no mundo da informática.

Por meio dessa lógica, o algoritmo computa a probabilidade de um usuário chegar a determinados sites ao clicar em *back links* aleatórios, logo, quanto maior o resultado desse cálculo probabilístico, maior a chance de ela aparecer na página principal do *Google*.

3.3 CÁLCULO DO PAGE RANK

Para saber o *Page Rank* de uma página são utilizados os conceitos da cadeia de Markov, que já foram explicitados anteriormente neste artigo. Para demonstrar esse processo, pode-se imaginar um diagrama de transições, como mostrado na Figura 5.

Figura 5 – Demonstração do *Page Rank* com a cadeia de Markov



Fonte: Autores.

Como exemplo, demonstrar-se-á, logo abaixo, a atuação do *Page Rank* em um microuniverso com somente 4 sites: o *G1*, indicado pelo número 1 na Figura 5 acima, o *Mercado-Livre*, pelo número 2, o *Youtube*, pelo número 3, e o *UOL*, pelo número 4.

Cada site do diagrama é um estado representado na Cadeia de Markov. As probabilidades de cada um, desses sites, recomendar outro podem ser vistas na Figura 5. Para o cálculo da classificação do *Page Rank* de qualquer página, é necessário organizar todas as probabilidades em uma matriz de transição. Os números 1, 2, 3 e 4, que representam cada site do microuniverso, também

representam sua posição na matriz, a exemplo de que o site G1 está localizado na linha e coluna 1 da matriz de transição de estado.

Tabela 1 – Matriz de transição

	G1	Mercado-Livre	Youtube	UOL
G1	0	1/3	1/3	1/3
Mercado-Livre	0	0	0	1
Youtube	1/2	0	0	1/2
UOL	0	0	1	0

Fonte: Autores.

As probabilidades apresentadas na Tabela 1 são as chances de ocorrer um determinado evento. Essa tabela mostra que uma transição do estado indicado na linha muda para o indicado na coluna. No caso do *Page Rank*, o evento é a chance de um usuário chegar a um determinado site por meio de cliques aleatórios nos *back links*. Um exemplo disto é, como dito na Tabela 1, que a chance de o internauta alcançar a página da *UOL*, a partir do site *G1*, dentre os *back links* é de um terço.

O algoritmo busca definir probabilidades estáticas de acesso aos sites por cliques aleatórios. Dessa maneira, a Cadeia de Markov utilizada será calculada com base no tempo discreto, um número tão grande a ponto de que essas probabilidades não se alterem mais; diferentemente de uma cadeia em tempo contínuo, que verificaria as probabilidades somente em um determinado momento.

A fim de definir a classificação de uma página, o algoritmo do *Page Rank* soma todas as probabilidades de um site ser recomendado.

Para os cálculos abaixo, os seguintes comentários serão necessários:

- O micro universo utilizado no exemplo trata-se de uma cadeia de Markov ergódica, que, a partir de qualquer um dos estados, consegue transitar-se para qualquer outro;
- P1 -> Representa a probabilidade que o site *G1* será acessado por outro;
- P2 -> Representa a probabilidade que o site *Mercado-Livre* será acessado por outro;
- P3 -> Representa a probabilidade que o site YouTube será acessado por outro;
- P4 -> Representa a probabilidade que o site UOL será acessado por outro;
- A soma de todas estas probabilidades (P1 + P2 + P3 + P4) sempre resultarão em 1.

Após os cálculos, considerando que o estado de equilíbrio será alcançado, as probabilidades P1, P2, P3 e P4 consequentemente representarão o *Page Rank* de cada uma das páginas. Assim, pode-se afirmar que o *Page Rank* de uma determinada página é a soma de todos os *Page Ranks* dos sites que a recomendaram.

- Vamos aos cálculos:

- *Page Rank* dos sites:
P1 = (1/2)*P3;
P2 = (1/3)*P1;
P3 = (1/3)*P1 + P4;
P4 = (1/3)*P1 + P2 + (1/2)*P3

- Cálculo do *Page Rank*

1 = P1 + P2 + P3 + P4
1 = (1/2)*P3 + (1/6)*P3 + P3 + (1/6)*P3 + (1/6)*P3 + (1/2)*P3
1 = (1/2)*P3 + (1/6)*P3 + P3 + (1/6)*P3 + (1/6)*P3 + (1/2)*P3
1 = (1/2)*P3 + (1/6)*P3 + P3 + (1/6)*P3 + (1/6)*P3 + (1/2)*P3
1 = 2*P3 + (1/2)*P3
1 = (5/2)*P3
P3 = 2/5 ou 0,400 (aproximadamente);
P1 = 2/10 ou 0,200 (aproximadamente);
P2 = 2/30 ou 0,066 (aproximadamente);
P4 = 2/6 ou 0,334 (aproximadamente);

Após os cálculos, pode ser concluído que:

- A página da internet com o maior *Page Rank* do exemplo é o *YouTube*, com a classificação de 0.400, e portanto, o site de maior relevância;
- A página da internet com o menor *Page Rank* do exemplo é o *Mercado-Livre*, com a classificação de 0.066, portanto, o site de menor relevância;
- Logo, em uma pesquisa, a ordem que estes sites apareceriam na página do Google seriam: 1º- *YouTube*, 2º- *UOL*, 3º- *G1* e 4º- *Mercado-Livre*.

Os resultados atingidos com P1, P2, P3 e P4 são as probabilidades estacionárias, ou seja, elas não mudarão com o acréscimo ou decréscimo de visitas às páginas do micro universo.

O que o algoritmo do Page Rank faz é, em sua essência, replicar esses cálculos milhares de vezes em um universo muito maior de sites para trazer ao usuário uma maior comodidade e melhor qualidade na busca pela informação.

5 CONSIDERAÇÕES FINAIS

O *Page Rank* é um algoritmo que utiliza conceitos da Probabilidade e da Álgebra Linear para fornecer ao usuário o melhor resultado. Neste artigo, explicou-se desde o seu surgimento até os métodos e cálculos que são utilizados para obter os resultados.

O algoritmo apóia-se nos conhecimentos sobre a Cadeia de Markov, ao usá-la para qualificar a relevância de cada página analisada. Explanou-se detalhadamente, por meio de uma linguagem didática, a forma que este método matemático auxilia na eficiência da busca.

Com a finalidade de mostrar o funcionamento do *Google*, ferramenta de busca presente na vida dos estudantes, e pessoas do mundo todo que estão conectadas à rede de computadores, as informações aqui postas esclareceram eventuais curiosidades, além de contribuir em conhecimento sobre esse motor de busca.

REFERÊNCIAS

ALVES, Rui; DELGADO, Catarina. **Processos estocásticos**. Set. 1997. Disponível em: <http://www.engenharia-puro.com.br/edwin/PO-II/Alves,%20Rui%20-%20Economia-Porto%20-%20Processo%20de%20Markov.pdf>. Acesso em: 17 maio 2016.

AMAZONAWS. Disponível em: <https://s3.amazonaws.com/qcon-assets-production/images/provas/24060/Imagem%20007.jpg>. Acesso em: 26 abr. 2016 às 23:46.

BRIN, Sergey Mihailovich; PAGE, Lawrence Edward. **The anatomy of a large-scale hypertextual web search engine**. Disponível em: <http://infolab.stanford.edu/~backrub/google.html>.Acesso em: 12 maio 2016.

FERNANDES JÚNIOR, Divaldo Portilho; VARGAS JÚNIOR, Valdivino. **Conceitos e simulação de cadeias de Markov**. Disponível em: <http://www.sbpcnet.org.br/livro/63ra/conpeex/pivic/trabalhos/DIVALDO_.PDF>. Acesso em: 17 maio 2016.

LANGVILLE, A.N.; MEYER, C.D. **Google's page rank and beyond**: the science of search engine rankings. New Jersey: Princeton University Press, 2011. 224p.

SANTOS, Reginaldo S. **Cadeias de Markov**. Disponível em: <http://www.mat.ufmg.br/~regi/gaalt/markov.pdf>. Acesso em: 14 maio 2016.

SCHEIDIES, Nick. Google Follows These 8 Simple Rules (and So Should You). **INCOME**. Disponível em: <http://www.incomediary.com/google-follows-these-8-simple-rules-and-so-should-you>. Acesso em: 17 maio 2016 às 21:44.

SOUZA, Benício José de. **Resumo de processos markovianos.** Disponível em: <http://www.pcs.usp.br/~lca/mkv.pdf>. Acesso em: 12 maio 2016.

WIKIPEDIA – A encicoplédia livre. Disponível em: <https://pt.wikipedia.org/wiki/Diagrama_de_transi%C3%A7%C3%A3o_de_estados>. Acesso: 26 abr. 2016 às 23:36.

Data do recebimento: 19 de setembro de 2016
Data da avaliação: 22 de setembro de 2016
Data de aceite: 13 de Outubro de 2016

1. Graduando em Ciências da Computação – UNIVERSIDADE TIRADENTES - UNIT.
E-mail: mv_vinicius10@hotmail.com
2. Graduando em Ciências da Computação – UNIVERSIDADE TIRADENTES - UNIT.
E-mail: joaootaviobf21@gmail.com
3. Graduando em Ciências da Computação – UNIVERSIDADE TIRADENTES - UNIT.
E-mail: efraimleite@gmail.com
4. Graduando em Ciências da Computação – UNIVERSIDADE TIRADENTES - UNIT.
E-mail: gabrielunitone@gmail.com