Google™ a matriz com o vector próprio de ouro

Saiba como calcular um vector próprio de uma matriz valeu ao Google um império de 25 mil milhões de dólares

Jorge Buescu *

Se o leitor utilizava regularmente a Internet no final dos anos 90, talvez se recorde dos "motores de busca" relativamente primitivos que se usavam na altura. Possivelmente ainda se lembrará de nomes como AltaVista, Lycos ou Yahoo! (que ainda existe); eram estes os grandes competidores por um mercado emergente e em crescimento exponencial. Uma pesquisa nas dezenas de milhões de páginas da Web feitas por um bom motor de busca pouparia imenso tempo e trabalho ao utilizador. Para o proprietário do motor de busca, por outro lado, ser o mais utilizado significava (e significa) receitas gigantescas em termos de publicidade.

Supondo ainda que o leitor utilizava a Web por volta de 1998 ou 1999, assistiu certamente por experiência pessoal a uma verdadeira revolução na forma de pesquisar a Web. Retratando a minha experiência pessoal, era um utilizador incondicional do AltaVista. Os resultados eram razoavelmente fiáveis e oferecia uma série de outras funcionalidades úteis (por exemplo, tradução automática). Algures no final de 1998, um colega disse-me para apontar o meu browser para **www.google.com**. E ocorreu um milagre.

O Google era (e é) um motor de busca com aspecto minimalista. Mas era de uma rapidez surpreendente – quase instantâneo. E, mais do que isso, era muito mais preciso do que qualquer um dos outros motores de busca da altura. Enquanto numa pesquisa típica com o AltaVista, digamos, das primeiras 20 respostas 7 ou 8 são pertinentes e as restantes nem tanto, e ao mesmo tempo podem não ser detectadas páginas muito relevantes, com o Google as pesquisas são extraordinariamente acutilantes. Nada de relevante parece ficar deixado de fora; e geralmente basta olhar para as primeiras 10 ou 20 respostas para encontrar os documentos mais importantes. O leitor que decida com base na sua própria experiência: a eficiência do Google parece magia!

O Google era tão mais avançado do que todos os seus outros rivais que rapidamente se impôs como o motor de busca WWW de referência. Alguns dos seus competidores desapareceram; outros demoraram meia dúzia de anos a recuperar o atraso, e agora são actores secundários. Hoje em dia, "fazer uma pesquisa na Web" é sinónimo de "ir ao Google". Os criadores do Google, Sergei Brin e Larry Page, na altura dois jovens alunos de Stanford, são multimilionários. O Google entrou em bolsa em 2004, com um valor de cerca de 25 mil milhões de dólares.

E tudo isto é devido, literalmente, a resultados matemáticos de Álgebra Linear ao nível do 1.º ano da Universidade. Aquilo que projectou o Google para a estratosfera informática foi a inovação tecnológica que consistiu simplesmente nisto: a construção de um algoritmo para fazer o *ranking* de páginas Web. Esse algoritmo implica a construção de uma matriz (que já é conhecida como a *matriz do Google*) e o cálculo (aproximado) do seu vector próprio principal. Tudo isto matérias ao alcance de um aluno de 1.º ano.

A história começa em 1997. Sergei Brin e Larry Page, dois estudantes de Doutoramento da Universidade de Stanford, consideram a evolução a longo prazo dos motores de busca da Web. A Web teria nessa altura cerca de uma centena de milhão de páginas, e a sua evolução seria previsivelmente explosiva; de facto, hoje, dez anos depois, tem pelas melhores estimativas trinta mil milhões de páginas.

Ora, independentemente da evolução do número de páginas, há uma variável que fica razoavelmente constante: a capacidade humana de dar atenção a uma busca. Ou seja, quer em 1997 quer em 2007 um ser humano consegue apenas absorver os primeiros 10, 20, ou talvez 30 primeiros resultados da busca. Mas não mais do que isso. Ninguém tem tempo ou paciência para ver se, escondido na posição 200, está uma página relevante.

Assim, concluem Brin e Page, a única forma de fazer com que os motores de busca acompanhem a evolução explosiva da Web é fazer com que eles próprios acompanhem a evolução desta, para que mostrem sempre primeiro os resultados mais significativos. Os motores de busca de 1997, para classificar a importância das páginas, usavam essencialmente comparação de conteúdos através de bases de dados gigantescas; com uma Web 300 vezes maior, como a de hoje, eles seriam completamente disfuncionais.

A proposta de Brin e Page foi pública e fez parte dos seus trabalhos de Doutoramento. Ainda hoje se pode encontrar o seu artigo "*The anatomy of a large-scale hypertextual Web search engine*" no servidor de artigos da Universidade de Stanford, em http://infolab.stanford.edu/pub/papers/google.pdf. A esta distância é uma leitura fascinante. E, tratando-se de um trabalho académico, é totalmente aberto. *O Google não tinha segredos industriais!*

O coração do Google é um algoritmo matemático chamado PageRank. Para compreendermos o que é o PageRank é importante sabermos as funções que um motor de busca deve desempenhar. São essencialmente três:

(1) Percorrer toda a Web, localizando todas as páginas com acesso Web.
(2) Indexar os dados recolhidos no passo (1).
(3) Classificar a importância de cada página na base de dados, de forma que, quando um utilizador realiza uma pergunta, as páginas mais importantes sejam apresentadas primeiro.

Os passos (1) e (2) são comuns a todos os motores de busca, e o seu crescimento acompanha a Web. O primeiro faz-se enviando para a Web pro-

gramas designados por *Web crawlers* e o segundo guardando a informação relevante em bases de dados. O passo crítico é o terceiro. Aquilo que Brin e Page afirmam é que a precisão do passo (3) tem de ser cada vez maior, porque o número total de páginas a que um ser humano pode dar atenção, ao contrário do tamanho da Web, não cresce. Por exemplo, ao pesquisar a entrada "matrix" encontrei 161.000.000 de respostas. É impossível percorrer sequer uma pequena fracção delas.

Eis, então, a proposta de Brin e Page para a classificação da importância das páginas Web – o famoso PageRank. A ideia base é simples: uma página é tanto mais importante quanto mais vezes for referida por *links* a partir de outras páginas. Assim, uma medida razoável da importância de uma página Web é o número de *links* que apontam para ela: uma página para a qual apontem 500 *links* é, provavelmente, 100 vezes mais importante do que uma página para a qual apontem 5 *links*. Assim, uma primeira tentativa, ainda grosseira, para fazer um *ranking* das páginas Web é dizer que a importância x_j de uma página P_j é igual ao número de *links* inversos que apontam para P_j.

A ideia fundamental de Brin e Page é, em vez de ser o motor de busca a julgar a importância das páginas com base no seu conteúdo, deixar a decisão da importância da página à própria Web, com base numa "votação democrática via links".

No entanto, este processo ainda é demasiado grosseiro para funcionar bem. Por exemplo, é vulnerável a manipulação: se eu criar uma página fictícia com 500 *links* a apontar para P_j estou a aumentar artificialmente, talvez com objectivos comerciais, o *ranking* de P_j. Para resolver este problema, o que fazemos é dar apenas "um voto" a cada página: isto é, se a página P_i tem n_i ***links***, então atribuímos a cada um desses *links* apenas o valor $1/n_i$. Assim, seria irrelevante a tal página fictícia ter 500 *links* para P_j ou apenas um.

Por outro lado, é claro que os *links*, tal como as páginas, não têm todos a mesma importância: um *link* vindo do Yahoo! tem de valer muito mais do que um *link* vindo, por exemplo, do blogue de um curioso – porque a página do Yahoo! é muito mais importante do que a do blogue do curioso. Assim, os próprios *links* têm de ser proporcionais à importância da página que os fornece.

Juntando estes dois princípios, chegamos à conclusão de que, sendo $x(P_i)$ a importância da página P_i, e designando por L_i o conjunto de todas as páginas Web que possuem links para P_i,

$$x(P_i) = \sum_{P_j \in L_i} \frac{x(P_j)}{n_j} \qquad (1)$$

A equação acima pode lembrar o leitor do problema da galinha e do ovo: para saber a importância de uma página, temos de saber a de todas as outras que têm *links* para ela. Na verdade, isso é natural: a equação (1) é um sistema de equações lineares – todas as variáveis dependem simultaneamente de todas as outras, pelo que só se consegue encontrar a solução resolvendo em simultâneo para todas as variáveis. Problema: trata-se de um sistema com... 30 mil milhões de variáveis (e todos os dias o número cresce)!

E aqui a Álgebra Linear elementar acorre em nosso auxílio. A matriz H_{ij} do sistema (1) é, pela forma de construção, uma matriz *estocástica por colunas*: todas as colunas têm soma 1. Isso implica, em particular, que o maior valor próprio é 1 e que o vector próprio correspondente será o que interessa para a solução do problema.

No entanto, o problema ainda não se encontra bem formulado do ponto de vista (como já foi afirmado) de "engenharia matemática". A matriz do sistema (1) é uma matriz de Markov, o que supõe que as transições entre páginas se dão aleatoriamente através de *links*. No entanto, muitas páginas não têm *links* (ficheiros pdf, imagens, etc.). Chama-se a estas páginas *nós pendurados*; se atingíssemos essas páginas aleatoriamente, nunca sairíamos de lá. A forma de o modelo os evitar é atribuir a cada nó pendurado uma probabilidade uniforme de navegação para fora do nó.

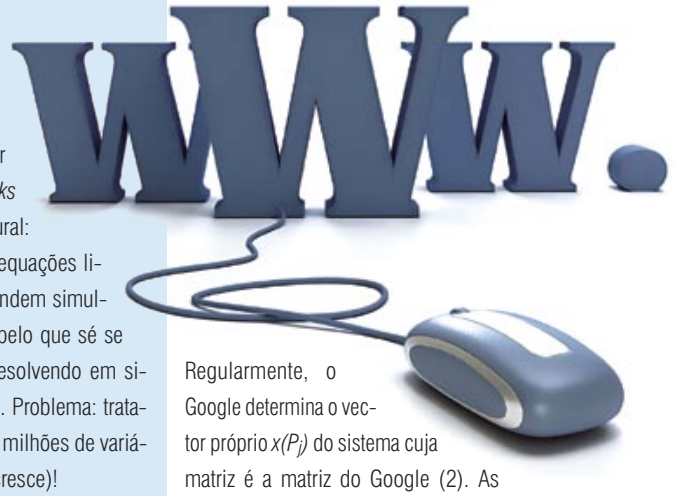
Este último ajustamento dá origem a um sistema com uma nova matriz,

$$G = \alpha H + (1-\alpha)\frac{1}{n} I \qquad (2)$$

onde H é a matriz dos *hyperlinks* do sistema (1), I é a matriz identidade e α é um parâmetro entre 0 e 1 e n é o número total de páginas.

G chama-se a *matriz do Google* e é ela que serve para definir as classificações de páginas, o famoso PageRank. Para $\alpha = 1$, G coincide com a antiga matriz H, e portanto dá o problema dos nós pendurados; para $\alpha = 0$ perde-se a informação sobre a estrutura da Web. Brin e Page chegaram à conclusão experimental de que o valor adequado é $\alpha = 0.85$.

E é esta a matriz que vale o seu peso em ouro. Um teorema clássico sobre matrizes (Perron-Frobenius, 1906) garante que existe um valor próprio dominante e que o vector próprio correspondente se pode calcular por métodos iterativos (o método da potência). Assim, o que o Google faz para calcular o PageRank é o seguinte.



Regularmente, o Google determina o vector próprio $x(P_j)$ do sistema cuja matriz é a matriz do Google (2). As componentes do vector próprio dão numericamente a importância $x(P_j)$ de cada uma das cerca de 30 mil milhões de páginas Web. A partir dos valores de $x(P_j)$ faz-se a ordenação – *ranking* – das páginas Web por ordem de importância absoluta. Este processo computacional é feito de uma vez por todas, e o *ranking* é guardado em base de dados. É este processo que tem o nome de PageRank.

Em resposta a uma pergunta que um utilizador faça ao Google – por exemplo, à minha pergunta sobre "matrix" – o Google vai pesquisar na sua base de dados as páginas que contenham esse texto. Em seguida, apresenta os resultados ordenados de acordo com o PageRank. Note-se que nesta fase o Google não faz cálculos nenhuns, mas apenas consulta de uma tabela: os cálculos do PageRank já foram efectuados anteriormente, à margem desta consulta.

Finalmente, cerca de uma vez por mês (ao que se diz) o Google refaz a determinação do PageRank. A razão é, evidentemente, acompanhar a evolução da própria Web, que durante um mês se transforma. É esta a razão pela qual a mesma pesquisa no Google, em dias consecutivos, pode dar resultados diferentes: significa que nesse intervalo o Google fez uma actualização do PageRank. Este fenómeno é conhecido por Google Dance.

E foi o PageRank que originou o império Google.

O Google, a sua matriz e os multi-milionários Sergei Brin e Larry Page são portadores de uma boa metáfora para o século XXI. Numa época em que cada vez mais se sobrevaloriza a informação e desvaloriza o conhecimento, é importante ter consciência de que uma boa ideia, desde que acompanhada do conhecimento científico apropriado – que até pode ser Álgebra Linear do 1.º ano – ainda pode mudar o Mundo. O Google mudou.

*** Professor na Faculdade de Ciências da Universidade de Lisboa**