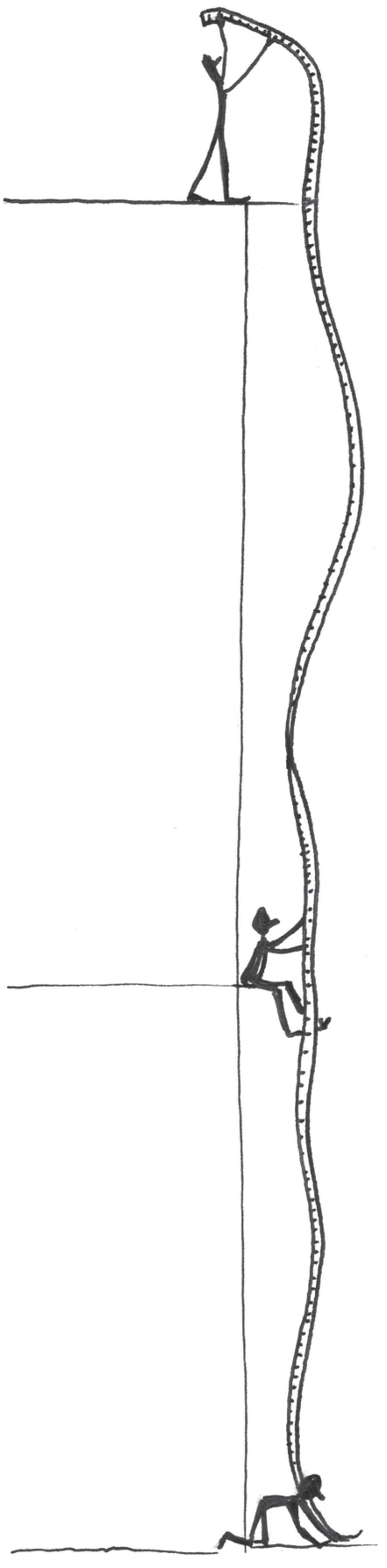
Miara ważności

*Krzysztof DIKS**

*Instytut Informatyki, Uniwersytet Warszawski



Czy zastanawiałeś się, drogi Czytelniku, dlaczego wyszukiwarka wyświetla adresy stron będących odpowiedzią na Twoje zapytanie właśnie w takiej, a nie innej kolejności? Zanim zacząłem pisać tę krótką notkę, zażądałem od wyszukiwarki `google.pl` odpowiedzi na pytanie „matematyka". Pierwszych pięć adresów do stron, które otrzymałem w odpowiedzi, to (1) `www.matematyka.org`, (2) `www.matematyka.pl`, (3) `www.math.edu.pl`, (4) `pl.wikipedia.org/wiki/Matematyka`, (5) `www.freewebs.com/podmatematyka`.

Google „pochwalił się", że wybrał je spośród 2 810 000 kandydatów. Dlaczego właśnie te uznano za najważniejsze? Jaka jest miara ważności strony? Okazuje się, że podstawą analizy ważności stron w Google jest analiza połączeń w grafie Internetu. Graf Internetu jest grafem skierowanym, w którym węzłami są strony, a krawędzie odpowiadają dowiązaniom pomiędzy stronami – strona zawierająca adres internetowy innej strony jest początkiem krawędzi, a strona o danym adresie jej końcem. Czy można na podstawie grafu Internetu powiedzieć, które strony są ważniejsze, a które mniej? Strona ważna to strona interesująca. Strona interesująca to taka, do której łatwo dotrzeć, ponieważ wiele innych stron na nią wskazuje. Na dodatek wiele spośród stron wskazujących na ważną stronę też jest ważnych i interesujących, itd. Larry Page i Sergey Brin, twórcy Google, wyrazili to matematycznie w następujący sposób.

Załóżmy, że mamy n stron $S_1, S_2, \ldots, S_n$. Niech $w(S_i)$ będzie liczbą rzeczywistą dodatnią mierzącą ważność strony S_i. Wówczas żądamy, żeby

$$w(S_i) = \sum_{S_j \in We(S_i)} \frac{w(S_j)}{|S_j|},$$

gdzie $We(S_i)$ to zbiór stron zawierających adres strony S_i (czyli zbiór początków krawędzi prowadzących do S_i), zaś $|S_j|$ jest liczbą odnośników (krawędzi) wychodzących ze strony S_j. Wzór ten oznacza, że ważność strony mierzy się ważnością stron na nią wskazujących. Jeżeli przyjmiemy na początek, że wszystkie strony są jednakowo ważne i dla każdej z nich $w(S_i) = \frac{1}{n}$, gdzie n to liczba wszystkich stron, to ważność stron można obliczyć za pomocą następującej metody iteracyjnej:

$$w_{k+1}(S_i) = \sum_{S_j \in We(S_i)} \frac{w_k(S_j)}{|S_j|}.$$

Na powyższy proces można spojrzeć jak na błądzenie losowe po grafie Internetu. Rozpoczynamy z dowolnej strony, a następnie z każdej oglądanej strony ruszamy losowo do jednej ze stron, do których adresy umieszczono na tej stronie. Jeśli nasz proces będziemy powtarzali bardzo długo, to pewne ze stron będziemy oglądali częściej niż inne. Strony oglądane częściej są ważniejsze.

Niestety, może się zdarzyć, że serfując po stronach, natrafimy na takie, z których nie ma wyjścia, np. strony zawierające zdjęcia. W takim przypadku zakładamy, że w następnym kroku wybierzemy losowo dowolną ze stron w Internecie. Teraz nasz proces iteracyjny wygląda następująco:

$$w_{k+1}(S_i) = \sum_{S_j \in We(S_i)} \frac{w_k(S_j)}{|S_j|} + \sum_{S_j \in K} \frac{w_k(S_j)}{n}$$

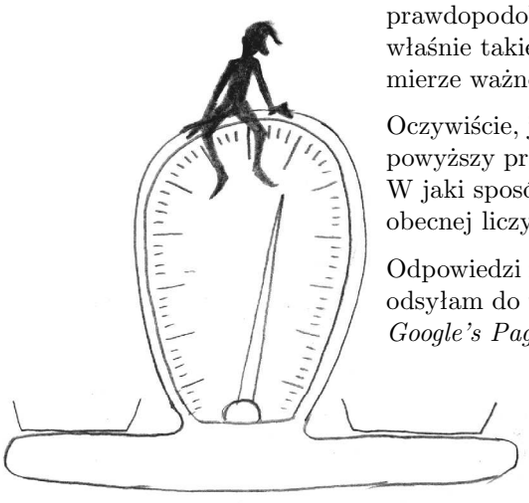
gdzie K oznacza zbiór wszystkich stron końcowych, czyli takich, które nie zawierają dowiązań do żadnych innych stron. Jesteśmy już blisko algorytmu Google ustalania miary ważności stron.

Brin i Page obserwując zachowanie użytkowników Internetu, zauważyli, że czasami porzucają oni bieżące przeszukiwanie i rozpoczynają nowe od (losowo) wybranej strony. Dlatego zmodyfikowali swój proces iteracyjny, wprowadzając parametr α o wartości z przedziału $(0, 1)$.

W każdej iteracji z prawdopodobieństwem α aktualne przeszukiwanie jest kontynuowane, natomiast z prawdopodobieństwem $1-\alpha$ rozpoczynane jest nowe przeszukiwanie. Ostatecznie postać każdej iteracji jest następująca:

$$w_{k+1}(S_i) = \alpha \left(\sum_{S_j \in We(S_i)} \frac{w_k(S_j)}{|S_j|} + \sum_{S_j \in K} \frac{w_k(S_j)}{n} \right) + (1-\alpha) \sum_{j=1}^{n} \frac{w_k(S_j)}{n}$$

Bardziej doświadczeni Czytelnicy pewnie już spostrzegli, że nasz proces iteracyjny jest procesem stochastycznym zbieżnym do stacjonarnego rozkładu prawdopodobieństwa. Celem kolejnych modyfikacji procesu było zapewnienie właśnie takiej zbieżności. Końcowy rozkład prawdopodobieństwa odpowiada mierze ważności stron.



Oczywiście, jest jeszcze wiele pytań, które pozostają bez odpowiedzi: Jak szybko powyższy proces zbiega do końcowego rozkładu? Jak dobierać parametr α? W jaki sposób przeprowadzać obliczenia na grafie Internetu, który w chwili obecnej liczy blisko 100 000 000 domen?

Odpowiedzi na te pytania to już inna historia. Wszystkich zainteresowanych odsyłam do wspaniałej książki autorstwa Amy N. Langville i Carla D. Meyera, *Google's PageRank and Beyond: The Science of Search Engine Rankings.*

Mebibajty i teraflopsy

Także w świecie komputerów posługujemy się różnymi miarami. Podstawową jednostką mierzącą ilość danych jest bajt (B). Jest to porcja danych złożona z 8 bitów (b), a więc mogąca przechować 2^8, czyli 256 różnych wartości. Także większe układy pamięciowe buduje się z cegiełek o wielkościach będących potęgami dwójki, a nie dziesiątki (nie znajdziemy komputera mającego 500 MB pamięci, za to 512 MB – owszem). Ponieważ jednak pewne potęgi liczb 2 i 10 różnią się bardzo nieznacznie (np. $2^{10} = 1024 \approx 1000 = 10^3$), więc przyjęło się używać przy mierzeniu pojemności tych samych przedrostków SI, które znamy z systemu dziesiętnego. I tak mówimy, na przykład, o 1 kB (kilobajcie), mając na myśli 1024 B, a nie 1000 B. Podobnie, 1 MB oznacza zwyczajowo $2^{20} = 1\,048\,576$ bajtów, czyli o $4,86\%$ więcej niż 10^6 bajtów. Przy przedrostku T (tera) różnica sięga prawie 10%.

Oczywiście, z punktu widzenia pedanta takie użycie przedrostków to błąd, dlatego powstał nowy zestaw przedrostków dla mnożnika 1024. I tak mamy kibibajty (1 KiB=1024 B), mebibajty (1 MiB=1024 KiB), dalej gibibajty (GiB), tebibajty (TiB), pebibajty (PiB)... Tak więc mój komputer ma 512 MiB, a nie 512 MB pamięci. Jak wszyscy wiemy z doświadczenia, ta nomenklatura póki co nie zyskała popularności i wygląda na to, że nadal w kontekstach związanych z komputerami królować będą przedrostki dziesiętne używane w roli binarnych.

Innym ważnym parametrem jest prędkość, którą podaje się w Hz. Jeśli, na przykład, procesor ma prędkość 1 GHz, to znaczy, że bramki logiczne w tym procesorze mogą zmieniać stan 10^9 razy na sekundę. Nie jest to bardzo miarodajna informacja, ponieważ różne instrukcje wykonują się w różnej liczbie cykli, a nowe technologie budowy procesorów (superskalarne, wektorowe, wielordzeniowe) pozwalają na upakowanie kilku operacji w jednym cyklu. Jeszcze mniej sensu ma ta jednostka w odniesieniu do farm obliczeniowych złożonych z wielu maszyn czy superkomputerów, wykonujących, na przykład, obliczenia z zakresu biomedycyny lub prognozowania pogody. Prędkość maszyn używanych do masowego przeprowadzania takich obliczeń mierzy się najczęściej we FLOPS-ach, czasem w MIPS-ach. Te skróty oznaczają, odpowiednio, FLoating-point Operations per Second, czyli liczbę wykonywanych operacji zmiennoprzecinkowych na sekundę, oraz Milion Instructions per Second, czyli miliony operacji na sekundę. Najszybsze superkomputery na świecie osiągają wydajność rzędu setek TFLOPS, czyli setek bilionów (10^{12}) operacji arytmetycznych na sekundę. Z kolei rozproszony projekt obliczeniowy Folding@Home (symulacje zwijania aminokwasów), w którym bierze udział około 250 000 procesorów w komputerach oraz – uwaga! – konsolach do gier z całego świata, przekroczył barierę 1 PFLOPS (petaFLOPS, 10^{15} operacji na sekundę). Nowoczesny komputer domowy ma w tych jednostkach moc rzędu kilku GFLOPS (10^9 operacji).

Michał ADAMASZEK