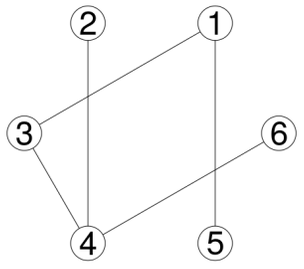
Google w łańcuchach

Łukasz RAJKOWSKI

Warto wspomnieć, że w kategorii „Słowo roku 2009” nominowany był czasownik *to fail*, który postąpił honorowo i przegrał. Jego porażka (sukces?) była jednak połowiczna, gdyż zwyciężył w kategorii „Najbardziej użyteczne słowo roku 2009”.

X_0 = Wstęp. W roku 2009 słowem dziesięciolecia stowarzyszenie *American Dialect Society* ogłosiło czasownik *to google*, którego polski odpowiednik – guglować/guglać – omawiany jest już na stronach Słownika Języka Polskiego, PWN. Nic dziwnego, wszak korzystanie z wyszukiwarki *Google* stało się elementem codzienności większości z nas i nie mamy skrupułów przed zadawaniem jej pytań o najbłahsze sprawy. Idąc za myślą przewodnią tego numeru *Delty*, o nieoczekiwanych związkach teorii z rzeczywistością, pokażemy, co wspólnego ma wszędobylska wyszukiwarka z teorią łańcuchów Markowa, sformułowaną w początkach XX wieku.



	1	2	3	4	5	6
%	20,25	9,93	20,04	29,75	10,10	9,92

X_1 = Przykład. Powiedzmy, że zwiedzamy graf spójny w taki sposób, że po znalezieniu się w danym wierzchołku, w kolejnym kroku odwiedzamy któregoś z jego sąsiadów, każdego z tym samym prawdopodobieństwem. Dla przykładu, kolejnymi odwiedzonymi wierzchołkami przy spacerowaniu po grafie na marginesie mogłyby być 1, 3, 4, 2, 4, 6, ... Tabela przedstawia procentowy czas spędzony w poszczególnych wierzchołkach w pewnej symulacji 10 000 kroków zwiedzania tego grafu. Zauważmy, że przedstawione wartości wydają się mocno powiązane ze stopniami (czyli liczbami sąsiadów) poszczególnych wierzchołków. Oczywiście nie jest to wcale przypadek, co niebawem wyjaśnimy.

X_2 = Internet. Załóżmy, że przeglądamy strony internetowe w bezmyślny sposób (nie róbcie tego w domu!), za każdym razem klikając w losowo wybrany odnośnik na stronie, na jakiej się znaleźliśmy. Gdyby procentowa liczba odwiedzin danej strony (na przykład `deltami.edu.pl`) stabilizowała się na pewnej granicznej wartości, to byłoby rozsądnie uznać tę wartość za *miarę ważności* tej strony – im jest większa ta wartość, tym częściej trafialibyśmy na daną stronę, „losowo” przeglądając Internet. Na pierwszy rzut oka nie wiadomo jednak, jak tę wartość obliczyć – pomogą nam w tym wspomniane we wstępie łańcuchy Markowa.

X_3 = Łańcuchy Markowa. *Łańcuchy Markowa* stanowią matematyczny model dla następującej (mało rzeczywistej, ale mam nadzieję dość obrazowej) sytuacji: wyobraźmy sobie układ N miast (ponumerowanych liczbami od 1 do N). W każdym z nich znajduje się teleport, który w losowy sposób wybiera miasto, do którego przenosi użytkownika. Niech teleport w mieście i przenosi do miasta j z prawdopodobieństwem p_{ij}. Załóżmy, że postanowiliśmy pozwiedzać miasta, codziennie korzystając (jednokrotnie) z zamieszczonych w nich teleportów.

Przyjmijmy, że miejsce startu („dzień zero”) jest losowe; z prawdopodobieństwem π_i jest to miasto i. Z jakim prawdopodobieństwem pierwszego dnia znajdziemy się w mieście j? Szansa na rozpoczęcie z k-tego miasta i przejście do j-tego wynosi $\pi_k p_{kj}$, a zatem prawdopodobieństwo odwiedzenia j-tego miasta pierwszego dnia to $\pi_j^{(1)} = \sum_{k=1}^{N} \pi_k p_{kj}$. Z rozkładu w „dniu zero” $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$ otrzymaliśmy rozkład w pierwszym dniu $\boldsymbol{\pi}^{(1)} = (\pi_1^{(1)}, \ldots, \pi_N^{(1)})$. W tym sensie prawdopodobieństwa $(p_{ij})_{i,j\leqslant N}$ określają nam przekształcenie P, które z jednego rozkładu tworzy inny.

Rozkładem nazwiemy dowolny ciąg N liczb nieujemnych, sumujących się do 1.

X_4 = Rozkład stacjonarny. Przy założeniu $(\star)$, przedstawionym na marginesie, r-ta iteracja przekształcenia P jest *przekształceniem zwężającym*, tzn. zmniejsza odległość między rozkładami o ustalony z góry czynnik. Aby to stwierdzenie miało sens, musimy doprecyzować odległość między rozkładami – u nas będzie to $|\boldsymbol{\pi} - \boldsymbol{\nu}| = \sum_{k=1}^{N} |\pi_k - \nu_k|$. W tej sytuacji z twierdzenia Banacha o punkcie stałym wynika, że istnieje punkt stały przekształcenia P, czyli taki rozkład $\tilde{\boldsymbol{\pi}}$ (nazywany *stacjonarnym*), że $P(\tilde{\boldsymbol{\pi}}) = \tilde{\boldsymbol{\pi}}$, tzn. prawdopodobieństwo wystartowania z i-tego miasta jest takie samo, jak prawdopodobieństwo znalezienia się w nim pierwszego (więc również drugiego, trzeciego, ...) dnia. Z twierdzenia Banacha

$(\star)$ istnieje takie r, że dla dowolnych $i, j \leqslant N$ z dodatnim prawdopodobieństwem możemy się przenieść z miasta i do miasta j po dokładnie r dniach.

O twierdzeniu Banacha można więcej poczytać w artykule Jarosława Górnickiego w Δ_{18}^{10}.

wynika również, że rozkład $\tilde{\boldsymbol{\pi}}$ może zostać uzyskany z dowolnego rozkładu π poprzez iterację P, tzn. ciąg $P^n(\boldsymbol{\pi})$ (gdzie $P^n = P \circ \ldots \circ P$) jest coraz lepszym przybliżeniem $\tilde{\boldsymbol{\pi}}$.

X_5 = Czas powrotu. Kiedy rozkład $\tilde{\boldsymbol{\pi}}$ istnieje, to (zakładając $(\star)$) ma pewną ciekawą własność – π_i jest równy odwrotności średniego czasu oczekiwania na powrót do i-tego wierzchołka, co teraz uzasadnimy. Załóżmy, że rozpoczęliśmy naszą wędrówkę, losując startowe miasto z rozkładu stacjonarnego. Pewnego dnia znaleźliśmy się w i-tym mieście i zastanawiamy się, kiedy doń wrócimy. Liczbę dni, jakie upłyną do pierwszego powrotu, oznaczmy jako τ_i – jest to pewna losowa wielkość. Niech μ_i będzie wartością oczekiwaną τ_i, tzn.

$$\begin{aligned}(1)\qquad \mu_i &= \mathbb{P}(\tau_i = 1) + 2\cdot\mathbb{P}(\tau_i = 2) + 3\cdot\mathbb{P}(\tau_i = 3) + \ldots = \\ &= \mathbb{P}(\tau_i \geqslant 1) + \mathbb{P}(\tau_i \geqslant 2) + \mathbb{P}(\tau_i \geqslant 3) + \ldots\end{aligned}$$

Przez $\mathbb{P}(A \mid B)$ oznaczamy prawdopodobieństwo zajścia zdarzenia A pod warunkiem zajścia zdarzenia B.

Niech X_k będzie numerem miasta odwiedzonego k-tego dnia i niech A_{kl} oznacza zdarzenie, że k-tego dnia byliśmy w mieście i, do którego nie wróciliśmy przez kolejnych l dni. Zauważmy, że $\mathbb{P}(\tau_i \geqslant k+1) = \mathbb{P}(A_{0k} \mid X_0 = i)$ oraz $\mathbb{P}(X_0 = i) = \tilde{\pi}_i$, zatem korzystając z (1), dostajemy

$$(2)\qquad \tilde{\pi}_i\mu_i = \mathbb{P}(A_{00}) + \mathbb{P}(A_{01}) + \mathbb{P}(A_{02}) + \mathbb{P}(A_{03}) + \ldots$$

Ponieważ wyszliśmy od rozkładu stacjonarnego, więc $\mathbb{P}(A_{kl})$ zależy tylko od l (co może wymagać chwili zastanowienia), zatem suma pierwszych $n+1$ składników w (2) to

$$(3)\qquad \mathbb{P}(A_{n0}) + \mathbb{P}(A_{(n-1)1}) + \ldots + \mathbb{P}(A_{0n}) = \mathbb{P}(\overbrace{A_{n0} \cup A_{(n-1)1} \cup \ldots \cup A_{0n}}^{B_n}),$$

gdyż zdarzenia $A_{k(n-k)}$ są rozłączne, ponadto ich suma B_n to zdarzenie, że w pierwszych n dniach odwiedziliśmy co najmniej raz i-te miasto. Z założenia $(\star)$ można wywnioskować, że $\mathbb{P}(B_n)$ jest zbieżne do 1, a zatem z (2) i (3) wynika $\tilde{\pi}_i\mu_i = 1$.

X_6 = Liczba odwiedzin. Na zakończenie teoretycznych rozważań zauważmy, że skoro wartość oczekiwana liczby dni potrzebnych na powrót do i-tego miasta wynosi μ_i, to średnio „raz na μ_i" kroków wracamy do i, zatem jeśli $D_n^{(i)}$ jest liczbą dni spędzonych w i-tym mieście podczas pierwszych n dni, to $D_n^{(i)}/n$ staje się coraz bliższe $1/\mu_i = \tilde{\pi}_i$. Rozumowanie to można nietrudno uściślić przy użyciu Mocnego Prawa Wielkich Liczb.

X_7 = Przykład wyjaśniony. Zauważmy, że nasz spacer po grafie spójnym odpowiada łańcuchowi Markowa o prawdopodobieństwie przejścia $p_{ij} = \deg(i)^{-1}\mathbb{1}_{i\sim j}$, gdzie $\deg(i)$ to stopień wierzchołka i, a $\mathbb{1}_{i\sim j}$ przyjmuje wartość 1, jeśli j jest sąsiadem i, oraz 0 w przeciwnym przypadku. Niech e będzie liczbą krawędzi. Łatwo uzasadnić, że suma stopni wierzchołków w grafie wynosi $2e$, a zatem wagi $\tilde{\pi}_i = \deg(i)/(2e)$ stanowią rozkład. Zwróćmy uwagę, że

$$\sum_{k=1}^{N} \tilde{\pi}_k p_{kj} = \sum_{k=1}^{N} \frac{\deg(k)}{2e}\cdot\frac{\mathbb{1}_{k\sim j}}{\deg(k)} = \frac{\sum_{k=1}^{N}\mathbb{1}_{k\sim j}}{2e} = \frac{\deg(j)}{2e} = \tilde{\pi}_j,$$

zatem $\tilde{\boldsymbol{\pi}}$ jest rozkładem stacjonarnym – uzasadnia to zaobserwowany wcześniej fenomen dotyczący średniego czasu przebywania w danym wierzchołku. Dość nieoczekiwanym stąd wnioskiem jest to, że średnia liczba odwiedzin jest w tym przypadku własnością lokalną i zależy tylko od stopnia wierzchołka i liczby krawędzi; bez znaczenia jest struktura pozostałej części grafu (jak również liczba wierzchołków).

Zauważmy, że losowe przeglądanie stron również możemy potraktować jako spacerowanie po grafie, lecz tym razem jest to graf skierowany.

X_8 = Google i PageRank. Wróćmy do losowego przeglądania stron. Oczywiście, ono również określa nam pewien łańcuch Markowa. Gdyby spełnione było założenie $(\star)$, to średnia liczba odwiedzin danej strony (czyli szukana *wartość strony*) zbiegałaby do wartości rozkładu stacjonarnego dla tej strony. Niestety podczas opisanego spacerowania po stronach internetowych moglibyśmy wpaść w pewne pułapki; możemy na przykład trafić na stronę bez żadnego odnośnika lub w inny sposób naruszyć $(\star)$. Celem rozwiązania tego problemu

wprowadza się niezerowe prawdopodobieństwo „teleportacji” – przed kliknięciem w dowolny odnośnik mamy szansę $(1-\alpha)$ (lub 1, jeśli na danej stronie nie ma odnośników) na przeskoczenie do losowo wybranej strony, każdej z jednakowym prawdopodobieństwem. W odróżnieniu od wcześniejszego przykładu, nie możemy jednak w prosty sposób wskazać rozkładu stacjonarnego, a dokładne rozwiązanie odpowiedniego układu równań nie wchodzi w grę ze względu na ogromną liczbę stron internetowych. Wiemy jednak, że zgodnie z twierdzeniem Banacha o punkcie stałym możemy przybliżyć rozkład stacjonarny poprzez rozpoczęcie od dowolnego rozkładu (np. $\pi_i^{(0)} = 1/N$) i powtarzanie

$$\pi_i^{(n+1)} = \sum_{j \to i} \frac{\alpha \cdot \pi_j^{(n)}}{\deg(j)} + \sum_{\deg(j)=0} \frac{\alpha \cdot \pi_j^{(n)}}{N} + \sum_{j=1}^{N} \frac{(1-\alpha) \cdot \pi_j^{(n)}}{N}, \tag{4}$$

gdzie pierwsza suma obejmuje wszystkie strony wskazujące na i, a druga wszystkie „strony puste”. Ten sposób mierzenia i obliczania ważności strony został zaproponowany przez Larry'ego Page'a i Sergeya Brina w 1997 roku, ochrzczony *PageRank* i wykorzystany w stworzonej przez nich wyszukiwarce Google jako istotny czynnik przy decydowaniu o kolejności wyświetlanych stron internetowych. I chociaż informacja o wartościach PageRanku nie jest już dostępna publicznie, wiele wskazuje na to, że ciągle odgrywa on niemałą rolę w sposobie, w jaki Google sortuje wyniki. Czytelnikom Zainteresowanym polecam samodzielne wyguglanie szczegółów.

O algorytmie PageRank z odrobinę innej perspektywy można przeczytać w artykule Krzysztofa Diksa w Δ^8_{08}. Zachęcamy do lektury!

Zadania

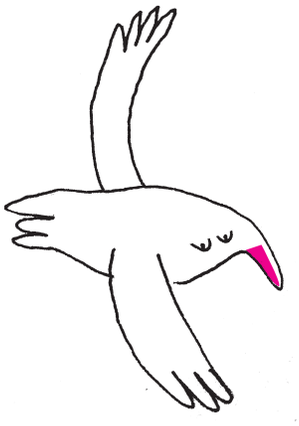
Przygotował Łukasz BOŻYK

M 1618. Do dyspozycji mamy n jednakowych świeczek. Pierwszego dnia zapalamy jedną świeczkę na dokładnie godzinę. Drugiego dnia wybieramy dwie świeczki i zapalamy je również na godzinę. Ogólnie k-tego dnia pewnych k świeczek pali się przez godzinę. Przypuśćmy, że n-tego dnia po upływie godziny wszystkie świeczki wypaliły się równocześnie. Wyznaczyć wszystkie wartości n, dla których taka sytuacja jest możliwa.
Rozwiązanie na str. 4

M 1619. Dwa rozłączne podzbiory zbioru $\{1, 2, \ldots, n\}$ mają tę samą sumę elementów. Wykazać, że każdy z tych podzbiorów ma mniej niż $n/\sqrt{2}$ elementów.
Rozwiązanie na str. 7

M 1620. Wyznaczyć wszystkie liczby całkowite $n \geqslant 2$ o następującej własności: Liczby całkowite od 1 do 16 można tak wpisać w pola tablicy 4×4, aby sumy liczb w wierszach i kolumnach były ośmioma parami różnymi liczbami całkowitymi, z których każda jest podzielna przez n.
Rozwiązanie na str. 15

Przygotował Andrzej MAJHOFER



F 987. Podczas rozładunku wagonów dębowe belki staczane są z pochylni wykonanej z dębowych desek. Jaka jest największa wartość kąta α między pochylnią i poziomem, dla której belki staczają się bez poślizgu? Należy przyjąć, że belki mają kształt walców, a współczynnik tarcia statycznego drzewa między powierzchniami z drzewa dębowego wynosi $f = 0{,}58$.
Rozwiązanie na str. 5

F 988. Dwa kółka (walce) o promieniach R i r $(R > r)$ wycięte z tego samego arkusza blachy osadzone są na dwóch równoległych osiach, wokół których mogą się swobodnie obracać. Większe z kółek wprawiono w ruch obrotowy, nadając mu prędkość kątową Ω_0, a następnie osie zsunięto tak, że kółka stykały się. Jaka ustaliła się końcowa prędkość kątowa większego z kółek? Przyjmij, że współczynnik tarcia kinetycznego między powierzchniami bocznymi kółek wynosi f i pomiń pozostałe opory ruchu (tarcie w zamocowaniu osi, opór powietrza itp.).
Rozwiązanie na str. 14