

Марковски вериги

и приложението им в Google

Евгения Стоименова

Институт по математика и информатика - БАН

август 2015, ЛИШ

<http://www.math.bas.bg/~jeni/jemarkov.html>

1

- I Случайно сърфиране
- II Движение по марковска верига
- III Дългосрочно поведение на марковска верига

2

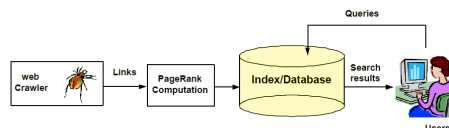
Случайно сърфиране

3

Търсене в интернет

Интернет търсачките като Google имат три основни действия:

- 1 Претърсват мрежата и локализируют всички страници с публичен достъп.
- 2 Индексират данните от стъпка 1, така че до тях да има ефективен достъп чрез ключови думи и изрази.
- 3 **Определят важността на всяка страница** в базата данни, така че по-важните страници да могат да се представят първи при произволна заявка.



В този курс ще разглеждаме стъпка 3.

4

Значимост на страниците

При търсене в интернет

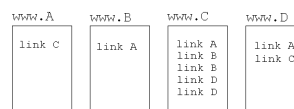
- Потребител задава **критерий** (ключови думи) за търсене в интернет мрежата World Wide Web.
- Естествено е да бъдат намерени **милиони** страници, свързани с критерия.

Идеята на Brin и Page (1998), Kleinberg (1999) е

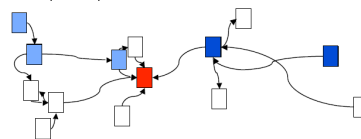
- Търсачката подрежда страниците съобразно тяхната значимост – обикновено това са страниците най-близки до критерия на търсене и най-полезни за потребителя.
- Да се дефинира значимостта на страниците посредством структурата на хипер-връзките в мрежата

5

Структура на хипер-връзките



WWW се разглежда като множество от страници, които са свързани помежду си чрез (хипер)връзки.



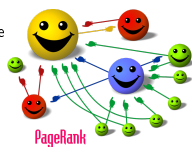
Към всяка страница са насочени връзки от други страници и от всяка страница излизат връзки към други страници от мрежата.

6

Как да се определя значимостта на страница?

Броят на връзките към една страница е показател за значимостта на страницата.

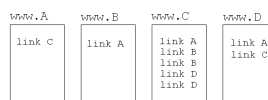
- Повече връзки към страницата – по-важна е страницата
- Връзка от "важна" страница има по-голямо тегло, отколкото връзка от "слаба" страница.
- Ако от страницата излизат няколко връзки, то теглото ѝ се разпределя пропорционално.
- Рангът на страницата A се определя от броя на страниците, които я цитират.



Как можем да знаем колко и какви страници сочат към всяка страница?

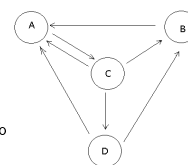
7

Сърфиране



WWW се представя като граф:

- Възлите на графа представляват страниците
- Движението (сърфирането) се извършва от един връх до друг само ако има стрелка в тази посока.



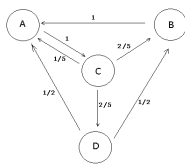
При продължително сърфиране потребителя посещава страниците с различна честота.

8

Случайно сърфиране

Ще изучаваме поведението на потребител, който се движи по мрежата като избира **случайно** страница, която да посети.

- Всеки път се избира връзка от списъка на страницата, в която се намира
- Смятаме, че всяка връзка от списъка е **равновероятно** да бъде избрана
- Преходът в нова страница зависи само от страницата, в която се намираме
- При продължително движение, се отчита честотата на посещения на страниците.



Като разделим честотата на посещение на една страница на общия брой на ходовете, получаваме оценка на вероятността страницата да бъде посетена.

Тази вероятност е оценка за важността на страницата.

9

PageRank на Google

Дефиниция

Рангът на една страница е отражение на поведението на потребителите на WWW и е равен на вероятността тя да бъде посетена при случайно сърфиране.

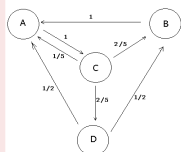
- Ранговете на страниците са свойство на страници, а не на начина за изчисляването.
- Един от начините за изчисляването на ранговете на страниците е чрез продължително случайно сърфиране.
- Математическият модел се описва с **марковска верига**.

10

Марковска верига

Марковска система

1. Системата има краен брой състояния.
2. Във всеки един момент тя се намира точно в едно от състоянията.
3. Преходът от едно състояние в друго се извършва при наличие на връзка и съгласно зададени вероятности.
4. Вероятностите за преход не се променят във времето.



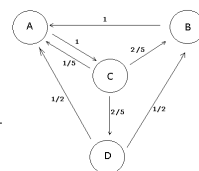
Редицата от последователни преходи на системата **A C D B A ...** образува марковска верига.

11

Марковска верига

Марковската система се определя от началното състояние и вероятностите за преход.

		към състояние			
		A	B	C	D
начално състояние	A	0	0	1	0
	B	1	0	0	0
	C	1/5	2/5	0	2/5
	D	1/2	1/2	0	0



Очевидно е, че моделът на случайното сърфиране е Марковска верига.

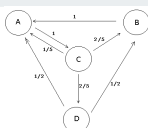
12

Марковска верига

Преходна матрица

- Матрицата от преходните вероятности е матрица, на която броя на редовете и стълбовете е равен на броя на състоянията на Марковската система.
- Първият ред на преходната матрица съдържа вероятностите за преход от състояние A до всяко състояние на системата.
- Вторият ред – вероятностите за преход от състояние B ...
- и т.н.
- **Сумата по редове е 1.**

$$\begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1/5 & 2/5 & 0 & 2/5 \\ 1/2 & 1/2 & 0 & 0 \end{pmatrix}$$



13

Начално състояние на системата

- Представя с вектор с дължина равна на броя на състоянията на системата, в който има една единица съответно на състоянието в което се намира системата и нули на всички останали позиции.

В примера: $q = (1 \ 0 \ 0 \ 0)$.

- Началното състояние може да се зададе и с вектор от вероятности,

$$q = (q_1 \ q_2 \ \dots \ q_k)$$

където q_i е вероятността системата да се намира в състояние i първоначално.

Дефиниция

q е **вероятностен вектор**, т.е. всички стойности са между 0 и 1 (включително) и $q_1 + q_2 + \dots + q_k = 1$.

14

Стохастична матрица

Дефиниция

Квадратна матрица с реални числа се нарича **стохастична**, ако

- (i) всички елементи на матрицата са между 0 и 1
- (ii) сумите на числата по редове е 1

Всеки ред на стохастичната матрица е вероятностен вектор.

Пример:

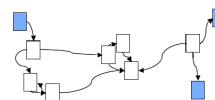
$$P = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1/5 & 2/5 & 0 & 2/5 \\ 1/2 & 1/2 & 0 & 0 \end{pmatrix}$$

15

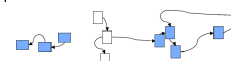
Случайно сърфиране

При това сърфиране има някои особености:

- Съществуват много страници, от които няма изходящи връзки.



- Има и такива страници, които създават затворен кръг, от които също не може да се излезе.

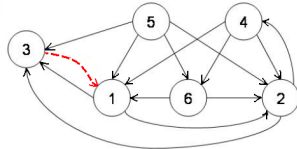


16

Страници без изходи

Пример "Прост капан".

$$P = \begin{pmatrix} 0 & 0.5 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0.33 & 0.33 & 0 & 0 & 0 & 0.33 \\ 0.25 & 0.25 & 0.25 & 0 & 0 & 0.25 \\ 0.5 & 0.5 & 0 & 0 & 0 & 0 \end{pmatrix}$$



Страниците без изход образуват "капани", които водят до поглъщане на ранга.

17

Първия модел - "Отегчен" сърфист

Да си представим, че сърфиста избира страница за преход случайно.

- Ако в страницата има няколко линка, то отегченият сърфист избира равновероятно една от тях.
- Ако в страницата няма линкове, сърфист отива на произволна, случайно избрана, страница в мрежата.

Правило 90 – 10

- В 90% от случаите потребителят избира случайно връзка от страницата, в която се намира. Всяка връзка е равновероятна.
- В останалите 10% от случаите той преминава към случайно избрана страница от мрежата, като при този преход всички страници от мрежата се смятат за равновероятни.

18

Преходна матрица при сърфиране по правилото 90-10

Вместо

$$\begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1/5 & 2/5 & 0 & 2/5 \\ 1/2 & 1/2 & 0 & 0 \end{pmatrix}$$

коригираме матрицата по правилото 90-10

$$\begin{pmatrix} 0.025 & 0.025 & 0.925 & 0.025 \\ 0.925 & 0.025 & 0.025 & 0.025 \\ 0.205 & 0.385 & 0.025 & 0.385 \\ 0.475 & 0.475 & 0.025 & 0.025 \end{pmatrix}$$

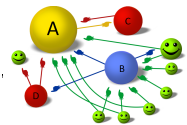
която всъщност се получава чрез

$$\begin{pmatrix} 0 & 0 & 0.90 & 0 \\ 0.9 & 0 & 0 & 0 \\ 0.18 & 0.36 & 0 & 0.36 \\ 0.45 & 0.45 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0.025 & 0.025 & 0.025 & 0.025 \\ 0.025 & 0.025 & 0.025 & 0.025 \\ 0.025 & 0.025 & 0.025 & 0.025 \\ 0.025 & 0.025 & 0.025 & 0.025 \end{pmatrix}$$

19

PageRank на Google

- Нека A е произволна страница.
- Всички страници, които имат връзка към A , допринасят за нейния ранг. Нека $M(A)$ е множеството от тези страници.
- Рангът на една страница u от $M(A)$ се разпределя към страница A и към другите страници, които са цитирани в u .
- Ако означим с $N(u)$ броя на връзките, които са в u , то приносът на u към ранга на A е $\frac{R(u)}{N(u)}$.



20

PageRank на Google

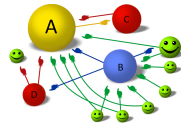
- Така рангът на страницата A се формира чрез ранговете на страниците, които имат връзки към нея:

$$R(A) = \frac{R(B)}{3} + R(C) + \frac{R(D)}{2} + \dots$$

- Чрез правилото 90-10

$$R(A) = c \sum_{u \in M(A)} \frac{R(u)}{N(u)} + (1-c)E(A),$$

където E е вектор с дължина броя на страниците в мрежата и елементи, съответни на някакви базови рангове на страниците (може и да са равни).



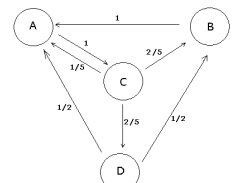
21

Симулация на марковска верига

Целта на симулираното движение по мрежата е да получим поредица от страници (номера), които да са посетени при случайното сърфиране.

- За симулацията имаме нужда от преходна матрица и начална страница.

	A	B	C	D
A	0	0	1	0
B	1	0	0	0
C	0.2	0.4	0	0.4
D	0.5	0.5	0	0



Очакваме при продължително сърфиране честотите на посещение на отделните страници да са много близки до ранговете им PageRank.

22

Симулация

- Симулацията ще направим с две функции – NextPage и Surfing.
- Входните аргументи на NextPage са текущата страница и преходната матрица, а изхода е следващата страница за един ход съгласно правилата на случайното сърфиране.
- Surfing прави зададен брой (T) прехода чрез NextPage и отчита броя на посещение за различните страници. Този брой се разделя на T и се получават честотите на посещение.

23

Симулация

Преход за една стъпка (NextPage)

Основната стъпка при симулацията на случайното сърфиране е преходът от една страница към друга за един ход.

- Променливата Page съдържа номера на текущата страница.
- Редът от преходната матрица $P[Page]$, съдържа преходните вероятности от страница Page до всичките N страници от мрежата.

	A	B	C	D
A	0	0	1	0
B	1	0	0	0
C	0.2	0.4	0	0.4
D	0.5	0.5	0	0

Когато знаем в коя страница е потребителят, задачата е да се генерира случайно цяло число (страница) между 0 и $N - 1$, съгласно зададените вероятности за тази страница.

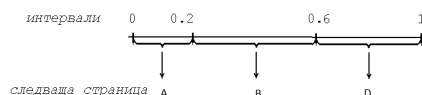
24

Симулация

Преход за една стъпка (NextPage)

	A	B	C	D
A	0	0	1	0
B	1	0	0	0
C	0.2	0.4	0	0.4
D	0.5	0.5	0	0

Използваме функция за генериране на случайно число в интервала $[0,1]$.



25

Симулация

Движение по мрежата (Surfing)

Движението по мрежата чрез случайно сърфиране има за цел да се отчетат честотите на посещение в различните страници.

- Функцията `Surfing` повтаря `NextPage` T на брой пъти.
- В масива `freq[]` се записва броя на посещенията във всяка страница.
- След T хода, елементът `freq[Page]` на масива ще съдържа честотата на поява на страницата `Page`.

26

Симулация

Движение по мрежата (Surfing)

- При зададен брой преходи T (напр. 1000), те се разпределят между страниците и се получават абсолютните честоти на посещение.
- Ако абсолютните честоти се разделят на общия брой (в случая 1000), се получават *относителните честоти* на посещение.
- Така тези честоти не са свързани повече с общия брой преходи, сумата им е 1, и се интерпретират по един и същ начин за различно по време сърфиране.

27

Експеримент

Започваме примерно от страница 0 и правим 1000 прехода

начална страница	Честоти на страниците			
	0	1	2	3
[0]	0.335	0.198	0.323	0.144
[0]	0.333	0.196	0.326	0.145
[0]	0.339	0.194	0.328	0.139
[0]	0.320	0.198	0.327	0.155
[0]	0.332	0.214	0.320	0.134
[0]	0.338	0.199	0.322	0.141

Крайните честоти от тези 6 експеримента са подобни.

28

Експеримент

Резултатите са подобни когато сърфирането започне от друга страница.

начална страница	Честоти на страниците			
	0	1	2	3
[1]	0.334	0.200	0.327	0.133
[1]	0.323	0.209	0.321	0.147
[1]	0.333	0.194	0.322	0.151
[2]	0.324	0.210	0.315	0.151
[2]	0.324	0.206	0.324	0.146
[2]	0.326	0.208	0.319	0.147
[3]	0.331	0.207	0.324	0.138
[3]	0.325	0.202	0.329	0.144
[3]	0.332	0.204	0.320	0.144

След 1000 прехода честотите на посещение на четирите страници не зависят от началната страница!

29

Експеримент

Честотите от отделните симулации са близки, но тяхната точност не е голяма. При 1000000 итерации честотите от една симулация са:

	Честоти на страниците			
страници	0	1	2	3
честоти	0.3310829	0.2045380	0.3231330	0.1412460

По-нататък ще разгледаме и едно друго приближение на честотите, следствие от случайното сърфиране, но осигуряващо по-бърза сходимост към истинските рангове. Чрез него могат да се получат следните честоти:

	Честоти на страниците			
страници	0	1	2	3
честоти	0.3257049	0.1848188	0.3181344	0.1713418

30

Действия с матрици

- Събиране на матрици

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} + \begin{pmatrix} 1 & -2 \\ 3 & 0 \end{pmatrix} = \begin{pmatrix} 1+1 & 2+(-2) \\ 3+0 & 4+0 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 3 & 4 \end{pmatrix}$$

- Умножение на матрица с число

$$5 * \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 5*1 & 5*2 \\ 5*3 & 5*4 \end{pmatrix} = \begin{pmatrix} 5 & 10 \\ 15 & 20 \end{pmatrix}$$

31

Действия с матрици

- Умножение вектор с матрица

$$\begin{pmatrix} 5 & 1 & 3 \end{pmatrix} * \begin{pmatrix} 5 & 1 \\ -2 & 0 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} 5*5 + 1*(-2) + 3*2 & 5*1 + 1*0 + 3*1 \end{pmatrix} = \begin{pmatrix} 36 & 8 \end{pmatrix}$$

(Дължината на вектора съвпада с броя на редовете на матрицата.)

$$\begin{pmatrix} 5 & 1 \\ -2 & 0 \\ 2 & 1 \end{pmatrix} * \begin{pmatrix} 1 \\ 3 \end{pmatrix} = \begin{pmatrix} 5*1 + 1*3 \\ (-2)*1 + 0*3 \\ 2*1 + 1*3 \end{pmatrix} = \begin{pmatrix} 8 \\ -2 \\ 5 \end{pmatrix}$$

(Броят на стълбовете на матрицата съвпада с дължината на вектора.)

32

Действия с матрици

► Умножение на матрици

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} * \begin{pmatrix} 3 & -1 & 2 \\ -2 & 0 & 2 \end{pmatrix}$$

$$= \begin{pmatrix} 1*3+2*(-2) & 1*(-1)+2*0 & 1*2+2*2 \\ 3*3+4*(-2) & 3*(-1)+4*0 & 3*2+4*2 \end{pmatrix} = \begin{pmatrix} -1 & -1 & 5 \\ 5 & -3 & 14 \end{pmatrix}$$

Броят на стълбовете на лявата матрица съвпада с броя на редовете на дясната матрица.

33

Движение по марковска верига

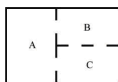


34

Пресмятане на вероятности

Пример: Обхождане на музея

- Нощен пазач обхожда музей по случаен начин и във всяка стая стои едно и също време.
- При преход от стая в стая, той избира изход равновероятно.

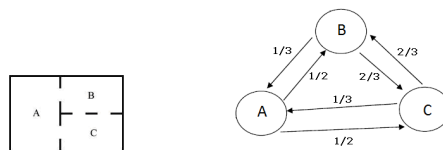


Каква е вероятността след 10 прехода пазачът да се окаже в първата стая?

Каква част от времето прекарва пазачът в различните стаи?

35

Описание на марковската верига



Очевидно редицата от последователни преходи образува Марковска верига.

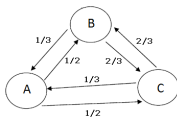
Преходна матрица:

		към		
		A	B	C
от	A	0	1/2	1/2
	B	1/3	0	2/3
	C	1/3	2/3	0

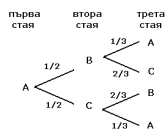
36

Условни вероятности

Нека първата стая е A. Намерете вероятностите след 2 прехода пазачът да се намира в стая A, B и C.



Пресмятаме вероятностите на различните пътища за 2 хода:



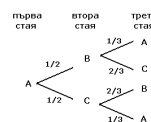
$$\begin{aligned} P(ABA) &= (1/2)(1/3) = 1/6 \\ P(ABC) &= (1/2)(2/3) = 1/3 \\ P(ACB) &= (1/2)(2/3) = 1/3 \\ P(ACA) &= (1/2)(1/3) = 1/6 \end{aligned}$$

37

Условни вероятности

Пресметнахме вероятностите на различните пътища за 2 хода:

$$\begin{aligned} P(ABA) &= 1/6 \\ P(ABC) &= 1/3 \\ P(ACB) &= 1/3 \\ P(ACA) &= 1/6 \end{aligned}$$



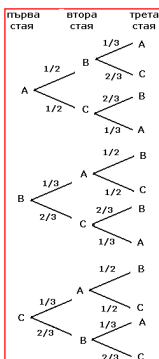
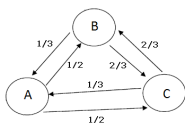
Условните вероятности за достигане на стаи A, B или C за 2 хода при условие, че началната стая е A са:

$$\begin{aligned} P(A|A) &= P(ABA) + P(ACA) = 1/6 + 1/6 = 1/3 \\ P(B|A) &= P(ABC) = 1/3 \\ P(C|A) &= P(ACB) = 1/3 \end{aligned}$$

38

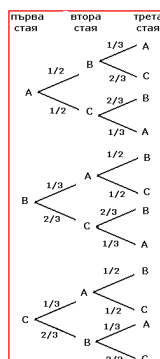
Дървовидна диаграма

Да се намерят вероятностите за преход до стаи A, B и C за 2 хода при различните начални позиции.



39

Дървовидна диаграма



$$\begin{aligned} P(ABA) &= 1/6 \\ P(ABC) &= 1/3 \\ P(ACB) &= 1/3 \\ P(ACA) &= 1/6 \end{aligned}$$

$$\begin{aligned} P(BAB) &= (1/3)(1/2) = 1/6 \\ P(BAC) &= (1/3)(1/2) = 1/6 \\ P(BCB) &= (2/3)(2/3) = 4/9 \\ P(BCA) &= (2/3)(1/3) = 2/9 \end{aligned}$$

$$\begin{aligned} P(CAB) &= (1/3)(1/2) = 1/6 \\ P(CAC) &= (1/3)(1/2) = 1/6 \\ P(CBA) &= (2/3)(1/3) = 2/9 \\ P(CBC) &= (2/3)(2/3) = 4/9 \end{aligned}$$

40

Вероятности при 2 хода

Пресметнахме

$$\begin{aligned} P(ABA) &= 1/6 \\ P(ABC) &= 1/3 \\ P(ACB) &= 1/3 \\ P(ACA) &= 1/6 \end{aligned}$$

$$\begin{aligned} P(BAB) &= 1/6 \\ P(BAC) &= 1/6 \\ P(BCB) &= 4/9 \\ P(BCA) &= 2/9 \end{aligned}$$

$$\begin{aligned} P(CAB) &= 1/6 \\ P(CAC) &= 1/6 \\ P(CBA) &= 2/9 \\ P(CBC) &= 4/9 \end{aligned}$$

Вероятности за достигане на стаи A, B или C за 2 хода при начална позиция A са:

$$\begin{aligned} P(A|A) &= P(ABA) + P(ACA) = 1/6 + 1/6 = 1/3 \\ P(B|A) &= P(ACB) = 1/3 \\ P(C|A) &= P(ABC) = 1/3 \end{aligned}$$

Вероятности за достигане на стаи A, B или C за 2 хода при начална позиция B са:

$$\begin{aligned} P(A|B) &= P(BCA) = 2/9 \\ P(B|B) &= P(BAB) + P(BCB) = 1/6 + 4/9 = 11/18 \\ P(C|B) &= P(BAC) = 1/6 \end{aligned}$$

Вероятности за достигане на стаи A, B или C за 2 хода при начална позиция C са:

$$\begin{aligned} P(A|C) &= P(CBA) = 2/9 \\ P(B|C) &= P(CAB) = 1/6 \\ P(C|C) &= P(CAC) + P(CBC) = 1/6 + 4/9 = 11/18 \end{aligned}$$

41

Преходна матрица за 2 хода

Пресметнахме условните вероятности:

$$\begin{aligned} P(A|A) &= 1/3 \\ P(B|A) &= 1/3 \\ P(C|A) &= 1/3 \end{aligned}$$

$$\begin{aligned} P(A|B) &= 2/9 \\ P(B|B) &= 11/18 \\ P(C|B) &= 1/6 \end{aligned}$$

$$\begin{aligned} P(A|C) &= 1/3 \\ P(B|C) &= 1/3 \\ P(C|C) &= 1/3 \end{aligned}$$

Построяваме преходната матрица за 2 хода:

		към		
		A	B	C
от	A	1/3	1/3	1/3
	B	2/9	11/18	1/6
	C	2/9	1/6	11/18

42

Преходна матрица за 2 хода

Ще покажем, че преходната матрица за 2 хода удовлетворява:

$$\begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 2/9 & 11/18 & 1/6 \\ 2/9 & 1/6 & 11/18 \end{pmatrix} = \begin{pmatrix} 0 & 1/2 & 1/2 \\ 1/3 & 0 & 2/3 \\ 1/3 & 2/3 & 0 \end{pmatrix} * \begin{pmatrix} 0 & 1/2 & 1/2 \\ 1/3 & 0 & 2/3 \\ 1/3 & 2/3 & 0 \end{pmatrix} = P^2$$

$$\text{Означихме: } P = \begin{pmatrix} p_{11} & p_{12} & \dots & p_{1k} \\ p_{21} & p_{22} & \dots & p_{2k} \\ \dots & \dots & \dots & \dots \\ p_{k1} & p_{k2} & \dots & p_{kk} \end{pmatrix}$$

Следователно

$$\begin{aligned} p_{31}(2) &= p_{31} \cdot p_{11} + p_{32} \cdot p_{12} + p_{33} \cdot p_{13} \\ &= (1/3) \cdot 0 + (2/3) \cdot (1/3) + 0 \cdot (1/3) \\ &= P(A|C) \end{aligned}$$

43

Стохастична матрица

Теорема

Ако P и Q са стохастични матрици от един и същи ред, то PQ е стохастична матрица от същи ред.

Пример:

$$\begin{pmatrix} 0 & 1/2 & 1/2 \\ 1/3 & 0 & 2/3 \\ 1/3 & 2/3 & 0 \end{pmatrix} * \begin{pmatrix} 0 & 1/2 & 1/2 \\ 1/3 & 0 & 2/3 \\ 1/3 & 2/3 & 0 \end{pmatrix} = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 2/9 & 11/18 & 1/6 \\ 2/9 & 1/6 & 11/18 \end{pmatrix}$$

Следствие

Ако P е стохастична матрица, то P^n е стохастична матрица.

Пример:

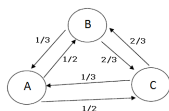
$$P^{20} = \begin{pmatrix} 0.25 & 0.3750000 & 0.3750000 \\ 0.25 & 0.3748998 & 0.3751002 \\ 0.25 & 0.3751002 & 0.3748998 \end{pmatrix}$$

44

Степени на преходната матрица

Да се намерят вероятностите за преход до стаи A, B или C за 3 и повече хода при различните начални позиции.

$$P = \begin{pmatrix} 0 & 1/2 & 1/2 \\ 1/3 & 0 & 2/3 \\ 1/3 & 2/3 & 0 \end{pmatrix}$$



$$P^3 = P^2 * P = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 2/9 & 11/18 & 1/6 \\ 2/9 & 1/6 & 11/18 \end{pmatrix} * \begin{pmatrix} 0 & 1/2 & 1/2 \\ 1/3 & 0 & 2/3 \\ 1/3 & 2/3 & 0 \end{pmatrix} = \begin{pmatrix} 0.22 & 0.39 & 0.39 \\ 0.26 & 0.22 & 0.52 \\ 0.26 & 0.52 & 0.22 \end{pmatrix}$$

$$P^5 = P^4 * P = \begin{pmatrix} 0.2469136 & 0.3765432 & 0.3765432 \\ 0.2510288 & 0.3086420 & 0.4403292 \\ 0.2510288 & 0.4403292 & 0.3086420 \end{pmatrix}$$

45

Степени на преходната матрица

$$P = \begin{pmatrix} 0 & 1/2 & 1/2 \\ 1/3 & 0 & 2/3 \\ 1/3 & 2/3 & 0 \end{pmatrix}$$

$$P^3 = \begin{pmatrix} 0.22 & 0.39 & 0.39 \\ 0.26 & 0.22 & 0.52 \\ 0.26 & 0.52 & 0.22 \end{pmatrix}$$

$$P^5 = P^4 * P = \begin{pmatrix} 0.2469136 & 0.3765432 & 0.3765432 \\ 0.2510288 & 0.3086420 & 0.4403292 \\ 0.2510288 & 0.4403292 & 0.3086420 \end{pmatrix}$$

$$P^{10} = \begin{pmatrix} 0.2499958 & 0.3750021 & 0.3750021 \\ 0.2500014 & 0.3692188 & 0.3807798 \\ 0.2500014 & 0.3807798 & 0.3692188 \end{pmatrix}$$

$$P^{20} = \begin{pmatrix} 0.25 & 0.3750000 & 0.3750000 \\ 0.25 & 0.3748998 & 0.3751002 \\ 0.25 & 0.3751002 & 0.3748998 \end{pmatrix}$$

46

Преход за n хода

Нека P е преходната матрица на Марковска верига.

$$P = \begin{pmatrix} p_{11} & p_{12} & \dots & p_{1k} \\ p_{21} & p_{22} & \dots & p_{2k} \\ \dots & \dots & \dots & \dots \\ p_{k1} & p_{k2} & \dots & p_{kk} \end{pmatrix}$$

Да означим с $p_{ij}(n)$ вероятността за преход от състояние i в състояние j за n прехода.

Теорема

Елементът (i, j) на матрицата P^n съдържа вероятността $p_{ij}(n)$.

$$P^n = \begin{pmatrix} 1 & \dots & j & \dots & k \\ \dots & \dots & \dots & \dots & \dots \\ i & \dots & p_{ij}(n) & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ k & \dots & p_{ki}(n) & \dots & p_{kk}(n) \end{pmatrix}$$

47

Преход за n хода

Какво се случва с преходната матрица, когато броя преходите расте неограничено?

В примера за музея получихме

$$P^{20} = \begin{pmatrix} 0.25 & 0.3750000 & 0.3750000 \\ 0.25 & 0.3748998 & 0.3751002 \\ 0.25 & 0.3751002 & 0.3748998 \end{pmatrix}$$

При голям брой преходи:

- Редовете на матрицата стават почти еднакви.
- Това означава, че вероятността за достигане на състояние j от състояние i е една и съща, независимо от началното състояние.
- С други думи, веригата "забравя" откъде е тръгнала, когато n е голямо. (това не е вярно за всяка верига)

48

Закон за големите числа

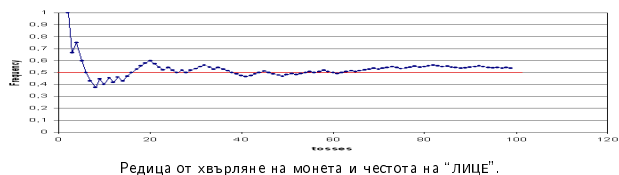
Преди 300 години в "Ars Conjectandi" е доказана най-известната вероятностна теорема.

Закон за големите числа

При голям брой независими повторения на експеримент, при който понякога настъпва дадено случайно събитие, честотата на появите на това събитие се приближава до неговата вероятност за настъпване.



Jakob Bernoulli
(1654-1705)



49

Закон за големите числа

- ▶ Закон за големите числа изисква независимост на събитията.
- ▶ За марковска верига редицата от последователни състояния не е от независими събития.
- ▶ За да приложим ЗГЧ трябва да си осигурим редица от независими събития

Закон за големите числа за марковска верига

Нека за първите n стъпки $H_j(n)$ е частта от тях, когато системата е в състояние s_j . Тогава

$$H_j(n) \rightarrow \pi_j$$

независимо от началното състояние.

- ▶ След определено време π_j е честотата на посещение на състоянието s_j . Така, независимо от началното състояние, системата ще се връща в състояние s_j безкрайно често.
- ▶ Рангът на една страница се дефинира чрез π_j .

50

Дългосрочно поведение на марковска верига



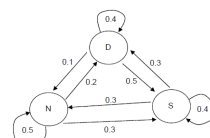
51

Пример: Таксиметрова компания

Таксиметрова компания извършва превози в 3 района: Northside, Downtown и Southside.

Компанията е установила, че:

- ▶ 50% от пътниците, качени в Northside, остават в Northside, 20% пътуват до Downtown, а 30% отиват до Southside.
- ▶ 10% от пътниците, качени в Downtown, пътуват до Northside, 40% остават в Downtown, а 50% отиват до Southside.
- ▶ 30% от пътниците, качени в Southside, пътуват до Northside, 30% пътуват до Downtown, а 40% остават в Southside.



$$\begin{pmatrix} 0.5 & 0.2 & 0.3 \\ 0.1 & 0.4 & 0.5 \\ 0.3 & 0.3 & 0.4 \end{pmatrix}$$

52

Преход за 1 ход

- ▶ Нека началното разпределение на такситата в града е: 20% в район Northside, 50% в Downtown, а 30% в Southside.

Така началният вектор е $q_0 = (0.20 \ 0.50 \ 0.30)$.

- ▶ Използвайки началното разпределение, можем да пресметнем процента на таксите във всеки регион след определен брой преходи.

- ▶ Например, да пресметнем колко таксите са в Downtown след 1 преход:

$$0.20P(ND) + 0.50P(DD) + 0.30P(SD) = 0.20(0.2) + 0.50(0.4) + 0.30(0.3) = 0.33.$$

Това число се получава и от произведението на вектора q_0 с втория ред на преходната матрица:

$$(0.20 \ 0.50 \ 0.30) \begin{pmatrix} 0.5 & 0.2 & 0.3 \\ 0.1 & 0.4 & 0.5 \\ 0.3 & 0.3 & 0.4 \end{pmatrix} = (\cdot \ 0.33 \cdot)$$

53

Състояние след 1 ход

Можем да намерим процента на такситата в Northside и Southside като умножим първия ред на P с q_0 и третия ред на P с q_0 , съответно.

Така разпределение на такситата след един преход се определя чрез:

$$q_1 = q_0 P = (0.20 \ 0.50 \ 0.30) \begin{pmatrix} 0.5 & 0.2 & 0.3 \\ 0.1 & 0.4 & 0.5 \\ 0.3 & 0.3 & 0.4 \end{pmatrix} = (0.24 \ 0.33 \ 0.43)$$

Следователно вероятността едно такси да е в Northside е 24%, вероятността да е в Downtown е 33% и вероятността да е в Southside е 43%.

54

Състояние след няколко хода

- ▶ Да намерим как са разпределени такситата след 2 прехода

$$\begin{aligned} q_2 &= q_1 P = (q_0 P) P = q_0 P^2 \\ &= (0.20 \ 0.50 \ 0.30) \begin{pmatrix} 0.36 & 0.27 & 0.37 \\ 0.24 & 0.33 & 0.43 \\ 0.3 & 0.3 & 0.4 \end{pmatrix} = (0.282 \ 0.309 \ 0.409). \end{aligned}$$

- ▶ Аналогично можем да намерим разпределението след 3 прехода:

$$q_3 = q_2 P = q_0 P P^2 = q_0 P^3.$$

- ▶ и след 4 прехода:

$$\begin{aligned} q_4 &= q_3 P = q_0 P^4 \\ &= (0.20 \ 0.50 \ 0.30) \begin{pmatrix} 0.3054 & 0.2973 & 0.3973 \\ 0.2946 & 0.3027 & 0.4027 \\ 0.3 & 0.3 & 0.4 \end{pmatrix} = (0.29838 \ 0.30081 \ 0.40081) \end{aligned}$$

Намираме разпределението след n прехода посредством уравнението

$$q_n = q_0 P^n$$

55

Дългосрочно поведение на марковската верига

- ▶ След 5 прехода: $q_5 = q_0 P^5 = (0.299514 \ 0.300243 \ 0.400243)$

- ▶ След 10 прехода: $q_{10} = q_0 P^{10} = (0.29999988 \ 0.30000059 \ 0.40000059)$

- ▶ След 20 прехода: $q_{20} = q_0 P^{20} = (0.3 \ 0.3 \ 0.4)$

- ▶ След 30 прехода: $q_{30} = q_0 P^{30} = (0.3 \ 0.3 \ 0.4)$

Изглежда, че векторът на състоянията клони към вектора

$$q_* = (0.3 \ 0.3 \ 0.4).$$

Този вектор наричаме **гранично разпределение**.

56

Стационарно разпределение на марковската верига

- При дадени преходна матрица P и начално състояние q_0 , граничното разпределение се получи чрез последователното пресмятане

$$q_{n+1} = q_n P = (q_{n-1} P) P = (q_{n-2} P) P^2 = \dots = q_0 P^n$$

като за достатъчно голямо n получихме

$$q_{n+1} = q_n \text{ и следователно } q_n = q_n P.$$

Когато броя на преходите расте неограничено (при $n \rightarrow \infty$) е възможно да достигнем стационарен вектор $\pi = (\pi_1 \dots \pi_k)$, за който е изпълнено

$$\pi = \pi P.$$

57

Стационарно разпределение на марковската верига

Откъде идва названието **стационарно разпределение**?

- Ако $\pi P = \pi$, то π е разпределението, което не се променя от преход на преход.
- В примера за таксиметрова компания: тъй като $q_* = (0.3 \ 0.3 \ 0.4)$ е стационарен вектор, ако 30% от такситата са в Northside, 30% са в Downtown и 40% са в Southside, то това съотношение ще бъде същото и след 1 ден.
- Всяко отделно такси може да се преместило от един район в друг, но след много превози (преходи) съотношението на таксите в районите ще остане постоянно.

58

Стационарно разпределение на марковската верига

- Елементите на стационарния вектор π , удовлетворява уравнението

$$\pi = \pi P,$$

и представляват вероятностите системата да е в различните състояния след продължителни преходи и **не зависи от началното състояние**.

- В примера за таксиметрова компания: всяко начално състояние ще доведе до едно и също разпределение на такситата след продължителни преходи.
- Тъй като различните начални състояния не влияят на стационарното разпределение, те може да влияят на времето за достигането му.

59

Намиране на стационарно разпределение

За стохастична матрица стационарно разпределение винаги съществува.

За да се намери стационарно разпределение π на една Марковска верига, с преходна матрица

$$P = \begin{pmatrix} p_{11} & \dots & p_{1k} \\ \dots & \dots & \dots \\ p_{k1} & \dots & p_{kk} \end{pmatrix}$$

трябва да се реши следната система от уравнения:

$$(\pi_1 \dots \pi_k) \begin{pmatrix} p_{11} & \dots & p_{1k} \\ \dots & \dots & \dots \\ p_{k1} & \dots & p_{kk} \end{pmatrix} = (\pi_1 \dots \pi_k),$$

където броя на неизвестните $(\pi_1 \dots \pi_k)$ е колкото са състоянията на Марковската система.

60

Намиране на стационарно разпределение

$$(\pi_1 \dots \pi_k) \begin{pmatrix} p_{11} & \dots & p_{1k} \\ \dots & \dots & \dots \\ p_{k1} & \dots & p_{kk} \end{pmatrix} = (\pi_1 \dots \pi_k)$$

Което означава да се реши системата уравнения:

$$\begin{cases} \pi_1 + \pi_2 + \dots + \pi_k = 1 \\ \pi_1 p_{11} + \pi_2 p_{21} + \dots + \pi_k p_{k1} = \pi_1 \\ \pi_1 p_{12} + \pi_2 p_{22} + \dots + \pi_k p_{k2} = \pi_2 \\ \dots \\ \pi_1 p_{1k} + \pi_2 p_{2k} + \dots + \pi_k p_{kk} = \pi_k. \end{cases}$$

61

Намиране на стационарно разпределение

В примера с таксиметровата компания:

$$\pi P = \pi$$

$$\pi \begin{pmatrix} 0.5 & 0.2 & 0.3 \\ 0.1 & 0.4 & 0.5 \\ 0.3 & 0.3 & 0.4 \end{pmatrix} = \pi$$

Уравненията са:

$$\begin{cases} \pi_1 + \pi_2 + \pi_3 = 1 \\ 0.5\pi_1 + 0.1\pi_2 + 0.3\pi_3 = \pi_1 \\ 0.2\pi_1 + 0.4\pi_2 + 0.3\pi_3 = \pi_2 \\ 0.3\pi_1 + 0.5\pi_2 + 0.4\pi_3 = \pi_3. \end{cases}$$

Заместваме $\pi_1 = 1 - \pi_2 - \pi_3$ в първите две уравнения и получаваме

$$\begin{cases} -0.5 + 0.6\pi_2 + 0.8\pi_3 = 0 \\ 0.2 - 0.6\pi_2 + 0.3\pi_3 = 0. \end{cases}$$

От тук следва

$$\pi_2 = 0.3 \quad \pi_3 = 0.4 \quad \pi_1 = 0.3$$

62

Регулярна марковска верига

- Винаги ли съществува **стационарно разпределение**?

Отговорът е: Да (Без доказателство на този етап.)

- В примера за таксиметрова компания: започнахме от разпределение q_0 и видяхме, че $q_0 P^n$ се приближава до граничен вектор π когато n става достатъчно голямо. Дали това е така винаги?

Ако марковската верига е регулярна, това е вярно.

63

Регулярна марковска верига

Кога можем да кажем, че една марковска верига клони към единствен стационарен вектор, независимо от началните условия?

Дефиниция

Една стохастична матрица се нарича **регулярна**, ако за някоя нейна степен нито един от елементи ѝ не е нула.

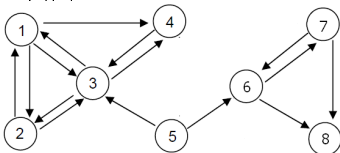
Дефиниция

Марковската верига се нарича **регулярна**, ако нейната преходна матрица е регулярна.

64

Нерегулярна марковска верига

Не всички марковски вериги са имат единствен стационарен вектор, към който клони редицата.



В зависимост от началното състояние, някои марковски вериги могат да имат различни стационарни вектори.

65

Единственост на стационарното разпределение

Теорема

Ако марковската верига е регулярна, то тя има само едно стационарно разпределение.

Преходната матрица на случайното сърфиране е регулярна, тъй като всичките ѝ елементи са положителни. Това беше направено по правилото 90-10 именно, за да си осигурим това свойство.

66

Примери за регулярни и нерегулярни марковски вериги

- Системата с преходна матрица $P = \begin{pmatrix} 0.8 & 0.2 \\ 0.1 & 0.9 \end{pmatrix}$ е регулярна, защото всичките ѝ елементи са положителни.
- Системата с преходна матрица $P = \begin{pmatrix} 0.15 & 0.85 \\ 1 & 0 \end{pmatrix}$ е регулярна, т.к.

$$P^2 = \begin{pmatrix} 0.1275 & 0.8725 \\ 0.15 & 0.85 \end{pmatrix}.$$

- Системата с преходна матрица $b = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ не е регулярна защото $P^2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, а $P^3 = P$. Преходна матрица се сменя между тези две матрици при следващи преходи. Следователно, тя винаги ще има нулеви елементи.

Системата има стационарен вектор $\begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$, но ако началното състояние е $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$, то $\begin{pmatrix} 1 \\ 0 \end{pmatrix} P = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ и $\begin{pmatrix} 1 \\ 0 \end{pmatrix} P^2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$. Системата се „върти“ между тези два вектора в зависимост от началното състояние и няма стационарен вектор.

67

Свойства на регулярна марковска верига

- За да е регулярна марковската верига, не е задължително всички елементи на преходната матрица да са ненулеви.
- След като всички стойности на матрицата са между 0 и 1 и строго положителни, то всички следващи степени на матрицата ще имат тези свойства и сходимостта е гарантирана.
- Последователните степени на преходната матрица образуват редица от матрици, която клони към една определена матрица P^* .
- Тази стационарна матрица има смисъл на преходна матрица за всички преходи от известно място нататък.

Теорема

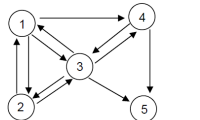
Ако марковската верига е регулярна, то стационарната матрица има едни и същи редове, които съвпадат със стационарното разпределение $(\pi_1 \dots \pi_k)$.

68

Поглъщаща марковска верига

Дефиниция

- Едно състояние на марковска верига се нарича поглъщащо, ако при достигане на това състояние системата остава там завинаги.
- С други думи вероятността да напуснем това състояние е 0.



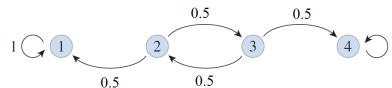
Дефиниция

Марковската верига се нарича поглъщаща, ако

- Има поне едно поглъщащо състояние
- За всяко състояние от системата вероятността да се достигне поглъщащо състояние за краен брой преход е по-голяма от 0.

69

Поглъщаща марковска верига



Неуверен комарджия играе на рулетка с \$20 по следната схема:

- На всяко завъртане на рулетката залага \$10 на червено.
- Ако излезе червено, той печели \$10, а ако е черно губи \$10.
- За простота нека приемем, че вероятността за червено е точно 1/2. (В действителност е малко по-малка)
- Човекът играе докато спечели \$30 или докато загуби всичките си пари. Тогава си тръгва.

Ще моделираме тази игра и ще изследваме дългосрочното поведение на веригата.

70

PageRank на Google

PageRank се дефинира по следния начин.

- Нека N е общия брой на web страниците в мрежата. Дефинираме матрицата на хипер-връзките:

$$Q_{ij} = \begin{cases} 1/N & \text{ако от страницата } i \text{ има връзка към страницата } j \\ 0 & \text{в останалите случаи,} \end{cases}$$

където N е общият брой на излизащите връзки от страницата i .

- За простота нека $Q_{ij} > 0$ за всички i . Това означава, че има връзка от страницата към себе си.
- Следователно Q може да се разглежда като преходна матрица на марковска верига на случайно сърфиране, където страниците представляват състоянията на марковската система.



71

PageRank на Google

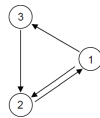
- Ако предположим, че веригата е регулярна, то съществува стационарно вероятностно разпределение $(\pi_1 \dots \pi_k)$ за посещение на състоянията (страниците).
- π_i е пропорционално на времето, което сърфиста посещава състоянието i . Колкото по-голямо е π_i , толкова по-значима е страницата.

Следователно PageRank на страницата i се дефинира с π_i .

72

PageRank Пример

Разглеждаме мрежа от 3 уеб страници: 1, 2, 3, такива, че
 $1 \rightarrow 1, 1 \rightarrow 2, 1 \rightarrow 3$
 $2 \rightarrow 1, 2 \rightarrow 2$
 $3 \rightarrow 2, 3 \rightarrow 3$.



- Преходната матрица на тази марковска верига е

$$Q = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 \end{pmatrix}.$$

- стационарният вектор на разпределението

$$\pi = (\pi_1 \quad \pi_2 \quad \pi_3)$$

удовлетворява уравненията

$$\pi = \pi Q \quad \text{и} \quad \pi_1 + \pi_2 + \pi_3 = 1.$$

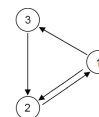
73

PageRank Пример

- Като решим системата линейни уравнения, получаваме:

$$(\pi_1 \quad \pi_2 \quad \pi_3) = \left(\frac{3}{9} \quad \frac{4}{9} \quad \frac{2}{9} \right).$$

- Следователно подреждането на страниците по ранг е:
 Webpage 2 > Webpage 1 > Webpage 3.



Интерпретация: Двете страници 1 и 3 сочат към 2 и следователно 2 е най-важна. Тъй като 2 сочи към 1, но не към 3, 1 е по-важна от 3.

74

Алгоритъм на PageRank

- Нека Q е преходна матрица на мрежа от N страници. Тъй като мрежата може да не е регулярна, я модифицираме:

$$P = \alpha \begin{pmatrix} Q_{11} & Q_{12} & \dots & Q_{1N} \\ Q_{21} & Q_{22} & \dots & Q_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ Q_{N1} & Q_{N2} & \dots & Q_{NN} \end{pmatrix} + \frac{(1-\alpha)}{N} \begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{pmatrix}$$

където $0 < \alpha < 1$.

- В този случай съществува единствен стационарен вектор.

Идея на алгоритъма:

- Страниците във всяка мрежа от N страници имат вътрешна значимост $(1-\alpha)/N$.
- Ако една страница i има значимост π_i , то тя отдава значимост $\alpha\pi_i$ като я разделя поравно между страниците, които цитира.

75

Алгоритъм на PageRank

Значимостта на страницата i се определя като решение на системата линейни уравнения:

$$(\pi_1 \quad \pi_2 \quad \dots \quad \pi_N) = \alpha (\pi_1 \quad \pi_2 \quad \dots \quad \pi_N) \begin{pmatrix} Q_{11} & Q_{12} & \dots & Q_{1N} \\ Q_{21} & Q_{22} & \dots & Q_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ Q_{N1} & Q_{N2} & \dots & Q_{NN} \end{pmatrix} + \frac{(1-\alpha)}{N} (1 \quad 1 \quad \dots \quad 1).$$

Тъй като

$$\sum_{i=1}^N \pi_i = 1,$$

това е еквивалентно на

$$(\pi_1 \quad \pi_2 \quad \dots \quad \pi_N) = (\pi_1 \quad \pi_2 \quad \dots \quad \pi_N) P.$$

76

PageRank на Google

- PageRank е измислен от
 - Sergey Brin and Lawrence Page (1998) The Anatomy of a Large-Scale Hypertextual Web Search Engine
 - Page L., Brin S., Motwani R. and Winograd T. (1998) The PageRank Citation Ranking: Bring Order to the Web, Technical Report, Stanford University.
- Larry Page и Sergey Brin са основателите на Google
- В сайта Google се намира следния текст:

'The heart of our software is PageRankTM, a system for ranking web pages developed by our founders Larry Page and Sergey Brin at Stanford University. And while we have dozens of engineers working to improve every aspect of Google on a daily basis, PageRank continues to provide the basis for all of our web search tools.'
- Официална информация на Google :
<https://sites.google.com/site/seooptimizers/>

77

Задачи

- Мрежа се състои от 5 web страници: A, B, C, D, E, със следните хипервръзки

$B \rightarrow A, C$
 $C \rightarrow B, B, D$
 $D \rightarrow C, E$
 $E \rightarrow A, B, D$

- Определете PageRank на страниците от мрежите. Постройте „Google матрица“, ако е необходимо да коригирате страници без изход.

78

Литература

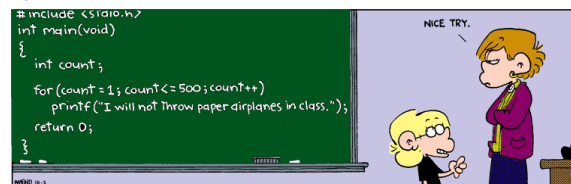
<http://www.math.bas.bg/~jeni/>

- Е. Стоименова (2013). Вериги на Марков и приложението им в Google. *Математика и информатика*, година LVI, кн.5, 2013, 426-443. (jemarkov2013p.pdf)
- Е. Стоименова (2012). Случайно сърфиране в Интернет. *Математика и информатика*, година LV, кн. 3, 225-237. (jesurf2012.pdf)
- Е. Стоименова (2013). Задачи за марковски вериги. (jeproblems2013.pdf)
- Waner, Costenoble. Finite mathematics, 2011. Chapter 7 Probability, Sec. 7.7 Markov systems. (Extract probability.pdf)
- Rebecca Atherton. A Look at Markov Chains and Their Use in Google. (Rebecca Atherton.pdf)
- Markov chains. (markov123.pdf)
- Grinstead, Snell. Introduction to probability. 1997. Chapter 11. (Chapter 11.pdf)
- Eileen Schoaff. Discrete Math projects. (Eileen11.txt)

79

Теми за проекти

- Алгоритми за подреждане в интернет търсачките
- Марковски вериги при оптимизация на светофари
- Симулации на пътен трафик
- Марковски процеси за вземане на решения
 - Priority Queuing Models for Hospital service
 - Design and Analysis of Hospital Admission Control
- Вероятностни парадокси
- Оценкаване на относителния дял чрез анкетни проучвания
- ...



80