Вериги на Марков и приложението им в Google

Евгения Стоименова
Институт по математика и информатика

Резюме. В тази статия се разглежда поведението на физически системи, които се развиват във времето. Моделите са стохастични и се представят с марковски вериги с краен брой състояния и постоянни вероятности за преход. Материалът е достъпен за ученици от средния курс. Базовите знания по вероятности от средния курс съществено помагат за разбирането на идеите. Използват се още алгебрични действия с вектори и матрици, които не затрудняват ученици с повишени интереси към информатиката. Текстът се основава на лекции, проведени от автора на Лятната изследователска школа на УчИМИ 2013.

1 Марковска верига, състояния, вероятности за преход

Веригите на Марков се използват за моделиране на последователни случайни събития, чието реализиране зависи от предишните настъпили събития. Да си представим една физическа система, която има n на брой състояния и във всеки един момент тя се намира само в едно от тези състояния. Системата преминава от едно състоя в друго по *случаен начин*. В n-тия момент на наблюдение системата се намира в състояние, които зависи от редицата от предишни състояния на системата. Да предположим, че състоянието в n-тия момент зависи само от това в кое състояние е била системата в предишния $n-1$ момент. Такава последователност от случайни събития образува марковска верига.

Исторически бележки. Веригите на Марков носят името на руския математик Андрей Марков (1856-1922), който първи започва да ги изу-

чава [1]. Прилагането им в интернет технологиите съществено повлия на развитието и структурата на интернет.

Пример 1. Нощен пазач обхожда музей по случаен начин. При преход от зала в зала, той избира изход равновероятно.

Да се представи графично движението на пазача и да се пресметнат вероятностите за преходи между залите.

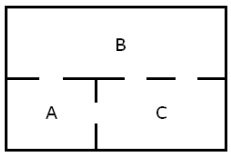


Схема на музея.

Да анализираме примера. Във всеки един момент пазачът се намира в едно от трите състояния: Пазачът е в състояние 1, ако се намира в зала A; в състояние 2, ако е в зала B и в състояние 3, ако се намира в зала C. (Номерирането не е от значение и тук е съответно на реда A-B-C). Ако пазачът се намира в състояние 1 в даден момент, то вероятността да премине в състояние 2 за едни ход е $1/2$ и в състояние 3 също $1/2$, тъй като има два равновероятни изхода от зала A. Записваме

$$p_{12} = 1/2, \qquad p_{13} = 1/2,$$

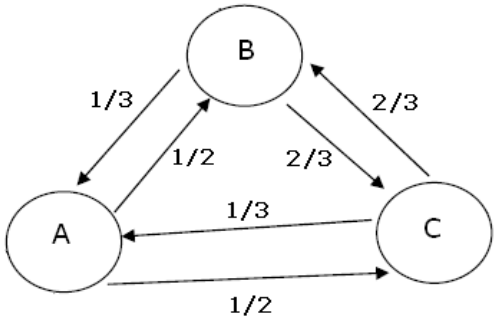
за да означим тези вероятности. С други думи с p_{ij} означаваме вероятността за преход от състояние i в състояние j за един ход.

Да пресметнем вероятностите за преход до останалите зали: Вероятността за преход от състояние 2 до състояние 1 е p_{21}=$1/3$; от състояние 2 до състояние 3 е $p_{23} = 2/3$; вероятността за преход от състояние 3 до състояние 1, p_{31}=$1/3$; от състояние 3 до състояние 2, $p_{32} = 2/3$. Представени в таблица, преходните вероятности са:

		към зала **A**	**B**	**C**
	A	0	$1/2$	$1/2$
от зала	**B**	$1/3$	0	$2/3$
	C	$1/3$	$2/3$	0,

където нулите съответстват на вероятност 0 за оставане в същата зала за един ход.

Можем да представим схемата на преходите и чрез *граф* по следния начин (Фиг. 1). Възлите на графа представляват залите (състоянията). Върховете са свързани, ако има преход между съответните зали. Стрелките означават посоките на преход и са надписани със съответните вероятности за преход.



Фигура 1: Схема на преходите в музея

1.1 Марковска верига

Дефиниция 1 *Марковска верига е случаен процес, представен като физическа система, която във всеки даден момент* $t = 1, 2, 3, \ldots$ *се намира в едно от краен (или изброим) брой състояния. На всяка стъпка, определена с* t*, системата преминава от едно състояние в друго по случаен начин, като вероятността за преход не зависи от* t*, а само от състоянието в което се намира системата.*

Вероятността p_{ij} за преход от състояние i в състояние j се нарича *преходна вероятност*.

Марковската верига се характеризира с *преходна матрица*, съдържаща преходните вероятност за всички състояния. Елементът (i, j) на матрицата съдържа вероятността p_{ij} за преход от състояние i в състояние j. Така всеки ред на матрицата съдържа вероятностите за преход от едно фиксирано състояние до всички състояния (включително настоящето) на системата.

$$\begin{pmatrix} p_{11} & p_{12} & \ldots & p_{1k} \\ p_{21} & p_{22} & \ldots & p_{1k} \\ \vdots & \vdots & \vdots & \vdots \\ p_{k1} & p_{k2} & \ldots & p_{kk} \end{pmatrix}. \tag{1}$$

Очевидно матрица е квадратна с размер $k \times k$, където k е броят на състоянията на системата.

Преходната матрица от примера с музея има вида:

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

Да отбележим някои от свойствата на преходната матрица:

(i) всички числа p_{ij} са между 0 и 1, включително,

(ii) сумите на числата по редове е 1,

(iii) за всяко j съществува i такова, че $p_{ij} > 0$.

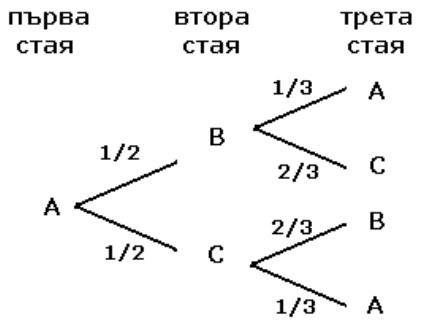
Марковската верига трябва да има *начално състояние*, с което да започне редицата. То се представя с вектор с дължина равна на броя на състоянията на системата, в който има една единица съответна на състоянието в което се намира системата и нули на всички останали позиции. В примера пазачът тръгва от състояние 1 (зала A), съответно началният вектор е $q = \begin{pmatrix} 1 & 0 & 0 & 0 \end{pmatrix}$.

В някои случаи веригата може да започне от различни състояния с определени вероятности. Началното състояние се задава с вектор от вероятности $q = \begin{pmatrix} q_1 & q_2 & \cdots & q_k \end{pmatrix}$, където q_i е вероятността системата да се намира в състояние i първоначално. Тъй като q е вероятностен вектор, всички стойности са между 0 и 1 и $q_1 + q_2 + \cdots + q_k = 1$. Този случай описва някои реални приложения и ще го разгледаме в Пример 2.

1.2 Преходни вероятности за два хода

Да направим някои пресмятания в примера. Първо да намерим вероятностите след 2 прехода пазачът да се намира в зала A, B или C. Да предположим, че първоначално пазачът се намира в зала A. Следвайки възможните преходи, след 1 ход той ще се намира в зала B с вероятност $1/2$ или в зала C също с вероятност $1/2$ (Фиг. 1). От зала B той може да продължи в зала A с вероятност $1/3$ или в зала C с вероятност $2/3$. Аналогично от зала C той може да продължи в зала A с вероятност $2/3$ или в зала A с вероятност $1/3$. Това последователно движение може да се изобрази с дървовидната диаграма на Фиг. 2.

Вероятностите за различните преходи от зала A намираме като умножим вероятностите по дървото:



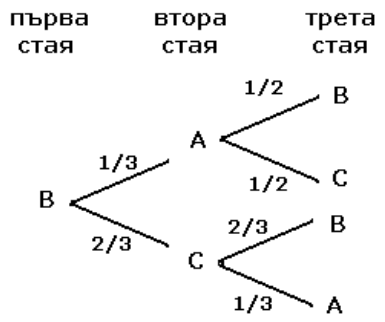
Фигура 2: Схема на вероятностите за преходи от зала A

$$P(A \to B \to A) = (1/2)(1/3) = 1/6$$
$$P(A \to B \to C) = (1/2)(2/3) = 1/3$$
$$P(A \to C \to B) = (1/2)(2/3) = 1/3$$
$$P(A \to C \to A) = (1/2)(1/3) = 1/6.$$

Сега да пресметнем вероятностите за преход до зали A, B или C за 2 хода при останалите начални позиции – от зала B и от зала C. Проследявайки дървото на фигурата в дясно, намираме последователно:

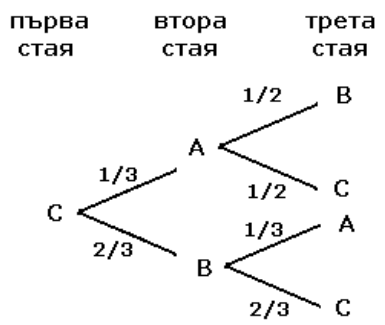
Тръгвайки от зала B,

$$P(B \to A \to B) = (1/3)(1/2) = 1/6$$
$$P(B \to A \to C) = (1/3)(1/2) = 1/6$$
$$P(B \to C \to B) = (2/3)(2/3) = 4/9$$
$$P(B \to C \to A) = (2/3)(1/3) = 2/9.$$



Тръгвайки от зала C,

$$P(C \to A \to B) = (1/3)(1/2) = 1/6$$
$$P(C \to A \to C) = (1/3)(1/2) = 1/6$$
$$P(C \to B \to A) = (2/3)(1/3) = 2/9$$
$$P(C \to B \to C) = (2/3)(2/3) = 4/9.$$



Така намираме, че вероятности за достигане на зали A, B или C за 2 хода при начална позиция A са:

$$\begin{aligned} P(A|A) &= P(A \to B \to A) + P(A \to C \to A) = 1/6 + 1/6 = 1/3 \\ P(B|A) &= P(A \to C \to B) = 1/3 \\ P(C|A) &= P(A \to B \to C) = 1/3. \end{aligned}$$

Вероятностите за достигане на зали A, B или C за 2 хода при начална позиция B са:

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

Вероятностите за достигане на зали A, B или C за 2 хода при начална позиция C са:

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

Дървовидните диаграми са удобни за пресмятане на условна вероятност, каквито са настоящите [8].

Да построим преходната матрица на веригата за 2 хода:

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

Ще покажем, че:

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

Ако означим с $p_{13}(2)$ вероятността за преход от състояние 1 в състояние 3 за 2 хода, то

$$\begin{aligned} p_{13}(2) &= p_{11} \cdot p_{13} + p_{12} . p_{23} + p_{13} \cdot p_{33} \\ &= (1/3) \cdot 0 + (2/3) \cdot (1/3) + 0 \cdot (1/3) \\ &= 2/9 = P(A|C), \end{aligned}$$

където числата p_{ij} са елементи от преходната матрица P за 1 ход, определена с (2). По правилото за умножение на матрици, елемента (1,3) на преходната матрица за 2 хода се е получил като произведение на третия ред на P с първия ѝ стълб.

Забележка: Използването на вектори и матрици не излиза от рамките на простите алгебрични действия събиране на матрици, умножение на матрица с число, умножение на две матрици. Съответно трябва да се имат предвид асоциативност при събиране и умножение и липсата на кумутативност при умножение на матрици.

1.3 Преходни вероятности за повече от два хода

Да видим как се определят вероятностите за преход до зали A, B или C за 3 и повече хода при различните начални позиции. Намираме ги като произведение на преходната матрица за два хода с преходната матрица за един ход[1]

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

[1]Тези и следващите умножения на матрици е за препоръчване да се правят с подходящи софтуерни функции.

С последователни умножения намираме следващите преходни матрици за 4, 5 и т.н. хода.

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

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

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

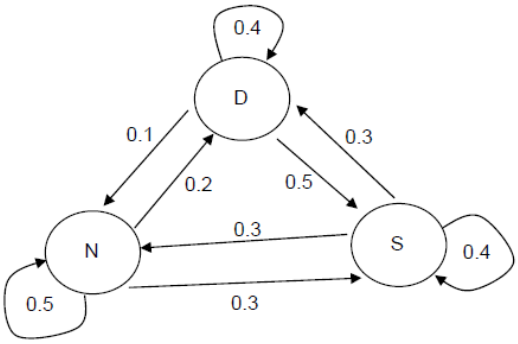
Теорема 1 *Елементът (i, j) на матрицата P^n съдържа вероятността $p_{ij}(n)$ за преход от състояние i в състояние j за n прехода.*

$$P^n = \begin{array}{c} \\ 1 \\ \dots \\ i \\ \dots \\ k \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & \dots & j & \dots & k \end{array} \\ \begin{pmatrix} p_{11}(n) & \dots & \dots & \dots & p_{1k}(n) \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & p_{ij}(n) & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ p_{k1}(n) & \dots & \dots & \dots & p_{kk}(n) \end{pmatrix} \end{array}$$

Какво се случва с преходната матрица, когато броя преходите расте неограничено? В примера за музея виждаме, че редовете на матрицата P^{20} са почти еднакви. Това означава, че при голям брой преходи 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. Добре е да знаем, че това не е вярно за всяка марковска система.

2 Състояние на марковска система след 1, 2 и повече хода

Ще разгледаме пример, в който повече от един обект извършват движение в марковска система.



Фигура 3: Схема на преходите на такситата

Пример 2. (Atherton [3]) Таксиметрова компания извършва превози в 3 района: Northside, Downtown и Southside. Копанията разполага с 3 паркинга за нощуване на такситата по един във всеки район. Компанията е установила, че:

- 50% от такситата, тръгващи сутрин от Northside, остават вечер в Northside, 20% пристигат в Downtown, а 30% пристигат в Southside.
- 10% от такситата, тръгващи сутрин от Downtown, пристигат вечер в Northside, 40% остават в Downtown, а 50% пристигат в Southside.
- 30% от такситата, тръгващи сутрин от Southside, пристигат вечер в Northside, 30% пристигат в Downtown, а 40% остават в Southside.

Компанията трябва да планира размерите на паркингите, така че да побират (без много излишна площ) такситата, които остават вечер във всеки един от трите района.

Решение: Нека началното разпределение на такситата в града е: 20% в район Northside, 50% в Downtown, a and 30% в Southside. Това означава, че вероятността случайно избрано такси да е в район Northside е 0.2, да е в район Downtown е 0.5 и да е в район Southside – 0.3.

Очевидно редицата от ежедневните преходи на едно такси образува марковска верига. Схемата на преходите на Фиг. 3 отразява условията на задачата. Преходните вероятности са записани в преходната матрица:

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

Началният (вероятностен) вектор на Марковската верига е

$$q_0 = \begin{pmatrix} 0.20 & 0.50 & 0.30 \end{pmatrix}.$$

Използвайки началното разпределение, ще пресметнем процента на такситата във всеки регион след определен брой хода.

2.1 Състояние след 1 ход

Като използваме началното разпределение и преходната матрица можем да намерим разпределението на такситата след един ход. Например да пресметнем колко таксита са в Downtown след 1 ход:

$$\begin{aligned} &0.20P(ND) + 0.50P(DD) + 0.30P(SD) \\ &= 0.20(0.2) + 0.50(0.4) + 0.30(0.3) = 0.33. \end{aligned}$$

Това означава, че след първия ден 33% от такситата са в Downtown. Това число се получава и от произведението на втория ред на преходната матрица с вектора q_0:

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

Можем да намерим процента на такситата в Northside и Southside като умножим q_0 с първия ред на P и q_0 с третия ред на P, съответно. Така разпределение на такситата след един ден, q_1, се определя чрез:

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

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

2.2 Състояние след няколко хода

Да намерим как са разпределени такситата след 2 хода.

$$\begin{aligned} q_2 &= q_1P = (q_0P)P = q_0P^2 \\ &= \begin{pmatrix}0.20 & 0.50 & 0.30\end{pmatrix} * \begin{pmatrix}0.36 & 0.27 & 0.37\\ 0.24 & 0.33 & 0.43\\ 0.3 & 0.3 & 0.4\end{pmatrix} \\ &= \begin{pmatrix}0.282 & 0.309 & 0.409\end{pmatrix}. \end{aligned}$$

Аналогично можем да намерим разпределението след 3 хода:

$$q_3 = q_2P = q_0PP^2 = q_0P^3$$

и след 4 прехода:

$$\begin{aligned} q_4 &= q_3P = q_0P^4 \\ &= \begin{pmatrix}0.20 & 0.50 & 0.30\end{pmatrix} * \begin{pmatrix}0.3054 & 0.2973 & 0.3973\\ 0.2946 & 0.3027 & 0.4027\\ 0.3 & 0.3 & 0.4\end{pmatrix} \\ &= \begin{pmatrix}0.29838 & 0.30081 & 0.40081\end{pmatrix} \end{aligned}$$

Намираме разпределението след n хода посредством уравнението

$$q_n = q_0P^n. \tag{3}$$

2.3 Дългосрочно поведение на марковската верига

Прилагайки формулата (3), получаваме разпределението на такситата след 5 хода:

$$q_5 = q_0P^5 = \begin{pmatrix}0.299514 & 0.300243 & 0.400243\end{pmatrix};$$

след 10 хода:

$$q_{10} = q_0P^{10} = \begin{pmatrix}0.299998819 & 0.3000005905 & 0.4000005905\end{pmatrix};$$

след 20 хода:

$$q_{20} = q_0P^{20} = \begin{pmatrix}0.3 & 0.3 & 0.4\end{pmatrix};$$

и след 30 хода:

$$q_{30} = q_0 P^{30} = \begin{pmatrix} 0.3 & 0.3 & 0.4 \end{pmatrix}.$$

Изглежда, че векторът на състоянията клони към вектора

$$q_* = \begin{pmatrix} 0.3 & 0.3 & 0.4 \end{pmatrix}.$$

Този вектор наричаме граничен вектор на разпределението или *гранично разпределение.*

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

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

като за достатъчно голямо n получихме

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

Когато броят на ходовете расте неограничено (при $n \to \infty$) е възможно да достигнем граничен вектор $\pi = \begin{pmatrix} \pi_1 & \dots & \pi_k \end{pmatrix}$, за който е изпълнено

$$\pi = \pi P.$$

Този граничен вектор е и *стационарен*, т.к. разпределението не се променя при умножение с преходната матрица. Това означава, че разпределението не се променя от преход на преход. Елементите на стационарния вектор представляват вероятностите системата да е в различните състояния след продължителни преходи и *не зависи от началното състояние.*

В примера за таксиметрова компания: тъй като $q_* = \begin{pmatrix} 0.3 & 0.3 & 0.4 \end{pmatrix}$ 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) съотношение в районите ще остане постоянно. Тъй като различните начални състояния не влияят на стационарното разпределение, те може да влияят на времето за достигането му.

2.4 Намиране на стационарно разпределение

Теорема 2 *За стохастична матрици стационарен вектор винаги съществува.*

Това твърдение оставяме без доказателство на този етап.

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 матричното уравнение $\pi P = \pi$, което записано в подробно е

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

където броя на неизвестните $\pi_1,\ \pi_2,\ \dots,\ \pi_k$ е колкото са състоянията на Марковската система. Това означава да се реши следната система от линейни уравнения:

$$\left|\begin{array}{rcl} \pi_1 + \pi_2 + \ldots + \pi_k & = & 1 \\ \pi_1 p_{11} + \pi_2 p_{21} + \ldots + \pi_k p_{k1} & = & \pi_1 \\ \pi_1 p_{12} + \pi_2 p_{22} + \ldots + \pi_k p_{k2} & = & \pi_2 \\ \ldots\ldots\ldots\ldots\ldots\ldots & \ldots & \ldots \\ \pi_1 p_{1k} + \pi_2 p_{2k} + \ldots + \pi_k p_{kk} & = & \pi_k. \end{array}\right.$$

В примера с таксиметровата компания стационарният вектор $\pi = \begin{pmatrix} \pi_1 & \pi_2 & \pi_3 \end{pmatrix}$ удовлетворява:

$$\begin{pmatrix} \pi_1 & \pi_2 & \pi_3 \end{pmatrix} * \begin{pmatrix} 0.5 & 0.2 & 0.3 \\ 0.1 & 0.4 & 0.5 \\ 0.3 & 0.3 & 0.4 \end{pmatrix} = \begin{pmatrix} \pi_1 & \pi_2 & \pi_3 \end{pmatrix}.$$

Това е еквивалентно на системата от линейни уравнения:

$$\left|\begin{array}{rcl} \pi_1 + s_2 + \pi_3 & = & 1 \\ 0.5\pi_1 + 0.1\pi_2 + 0.3\pi_3 & = & \pi_1 \\ 0.2\pi_1 + 0.4\pi_2 + 0.3\pi_3 & = & \pi_2 \\ 0.3\pi_1 + 0.5\pi_2 + 0.4\pi_3 & = & \pi_3. \end{array}\right.$$

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

$$\left|\begin{array}{rcl} -0.5 + 0.6\pi_2 + 0.8\pi_3 & = & 0 \\ 0.2 - 0.6\pi_2 + 0.3\pi_3 & = & 0 \end{array}\right.$$

и получаваме

$$\pi_2 = 0.3, \qquad \pi_3 = 0.4, \qquad \pi_1 = 0.3.$$

Забележка: Прави впечатление, че в системата за стационарния вектор, уравненията са с едно повече от неизвестните, имаме допълнително условие за сумата на елементите да дава 1. Това не преопределя системата, т.к. (може да се види в примера) едно от уравненията се получава като комбинация на другите две.

3 Регулярна марковска верига

Възникват естествени въпроси за съществуване и единственост на стационарното разпределение.

1. *Винаги ли съществува стационарно разпределение?*

   Отговорът е „Да" и ще го оставим без доказателството на този етап.

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

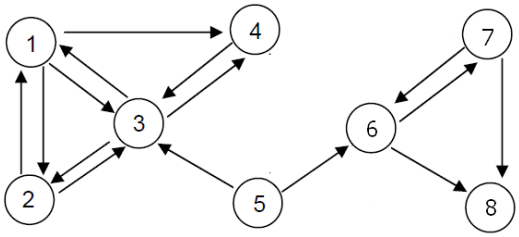
   Отговорът е „Това е вярно, ако марковската верига е регулярна". (Дефиницията е по-долу.)

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

Дефиниция 2 *Една стохастична матрица се нарича регулярна, ако за някоя нейна степен нито един от елементи ѝ не е нула.*

Да отбележим, че 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.

Дефиниция 3 *Марковската верига се нарича регулярна, ако нейната преходна матрица е регулярна.*



Фигура 4: Нерегулярна марковска верига

Не всички марковски вериги са такива, че имат единствен стационарен вектор, към който клони редицата. В зависимост от началното състояние, някои марковски вериги могат да имат различни стационарни вектори. Това се вижда на примера на Фиг. 4. Ако веригата започне от някое от състоянията 1, 2, 3 или 4, тя никога няма да достигне останалите състояния и обратно, ако веригата започне от някое от състоянията 6, 7 или 8, тя рано или късно ще завърши в състояние 8. Такива вериги не са регулярни.

Теорема 3 *Ако марковската верига е регулярна, то тя има само едно стационарно разпределение.*

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 4 *Ако марковската верига е регулярна, то стационарната матрица има едни и същи редове, които съвпадат с стационарно разпределение* $\begin{pmatrix}\pi_1 & \dots & \pi_k\end{pmatrix}$.

Няма да разглеждаме нерегулярни марковски вериги, въпреки че те са многобройни и със значими приложения през миналия век. Ще спомена само поглъщащата марковска верига. Такива верига има едно или повече поглъщащи състояния. При достигане на такова състояние, системата остава там завинаги (Фиг. 4, състояние 8). С други думи вероятността да напуснем това състояние е 0.

Тук ще разгледаме най-значимото съвременно приложение на регулярните марковските вериги – подреждането на списъка от намерените резултати на една интернет търсачка.

4 PageRank на Google

Създателите на интернет търсачките трябва да удовлетворят три основни групи клиенти: търсещите информация (които искат полезни резултати); рекламодателите (които искат връзка към рекламата си); и доставчиците на страници с рекламно място (които искат да максимизират приходите от реклама). Редът, в който се съставя списъка от намерените страници е съществена част при проектирането на една интернет търсачка.

Алгоритъмът на Google подрежда списъка от намерените страници съобразно вероятностите те да бъдат посетени от потребител, който „сърфира" в мрежата като избира по случаен начин хипер-връзка от страницата, в която се намира. Значимостта (рангът) на една страница е отражение на поведението на потребителите на WWW и е равен на вероятността тя да се бъде посетена при случайно сърфиране. Един начин за изчисляване на ранговете на уеб страници е описан в [2].

Да анализираме условията на задачата. Нека броя на всички страници в мрежата е N. От един връх i към връх j се задава стрелка, ако в страницата i има хипер-връзка към страницата j. Движението (сърфирането) се извършва от един връх до друг само ако има стрелка в тази посока. Да си представим, че сърфистът избира с равна вероятност една от всичките връзки в страницата, в която се намира. Дефинираме матрицата на хипер-връзките:

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

където $L(i)$ е общият брой на излизащите връзки от страницата i.

За простота нека $Q_{ii} > 0$ 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. Ако предположим, че веригата е регулярна, то съществува стационарно вероятностно разпределение $(\pi_1, \pi_2, \ldots, \pi_N)$ за посещение на състоянията (страниците). Така π_i е пропорционално на времето, което сърфистът посещава състоянието i. Колкото по-голямо е π_i, толкова по-значима е страницата. Следователно значимостта на страницата i се дефинира с π_i.

Реалната уеб мрежа не е регулярна и е възможно сърфистът да попадне в страница, от която не излизат връзки. За да се избегнат такива ситуации, преходната матрица се модифицира по следния начин:

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

където $0 < \alpha < 1$.

С други думи, сърфистът с вероятност $\alpha/L(i)$ избира една от връзките в страницата, в която се намира, ако $L(i) > 0$ и с вероятност $(1-\alpha)/N$ преминава в произволна страница от мрежата. Идеята на алгоритъма е, че: (i) страниците във всяка мрежа от N страници имат вътрешна значимост $(1-\alpha)/N$; (ii) ако една страница i има значимост π_i, то тя отдава значимост $\alpha\pi_i$ като я разделя поравно между страниците, които цитира.

За отбележим, че преходната матрица на случайното сърфиране е регулярна, тъй като всичките и елементи са положителни. Множителите α и $(1-\alpha)$ са необходими, за да си осигурим това свойство. За регулярна марковска верига съществува единствен стационарен вектор, удовлетворяващ $\pi P = \pi$. Стационарните вероятности показват каква част от времето прекарва сърфистът в различните страници. Така, ако $\pi_i > \pi_j$, то страницата i е по-значима от страницата j и нейния ранг трябва да е по-висок.

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

$$\begin{pmatrix} \pi_1 & \pi_2 & \ldots & \pi_N \end{pmatrix} = \alpha \begin{pmatrix} \pi_1 & \pi_2 & \ldots & \pi_N \end{pmatrix} \begin{pmatrix} Q_{11} & Q_{12} & \cdots & Q_{1N} \\ Q_{21} & Q_{22} & \cdots & Q_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ Q_{N1} & Q_{N2} & \cdots & Q_{NN} \end{pmatrix}$$

$$+\frac{(1-\alpha)}{N} \begin{pmatrix} 1 & 1 & \ldots & 1 \end{pmatrix}.$$

Тъй като

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

това е еквивалентно на

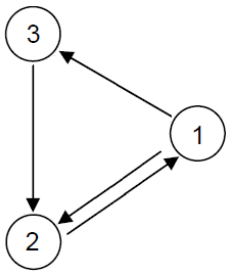
$$\begin{pmatrix}\pi_1 & \pi_2 & \dots & \pi_N\end{pmatrix} = \begin{pmatrix}\pi_1 & \pi_2 & \dots & \pi_N\end{pmatrix} P.$$

Решението на системата $\pi P = \pi$ е същността на алгоритъма за подреждане на Google, наречен PageRank. На теория не е толкова трудно да се оцени π, тъй като P^n клони много бързо към своята граница

$$\Pi = \pi 1 = \begin{pmatrix}\pi_1 & \pi_2 & \dots & \pi_N \\ \pi_1 & \pi_2 & \dots & \pi_N \\ \vdots & \vdots & \vdots & \vdots \\ \pi_1 & \pi_2 & \dots & \pi_N\end{pmatrix}.$$

Пример 3. Разглеждаме мрежа от 3 уеб страници, 1, 2 и 3, със следните хипервръзки:

$1 \to 1$, $1 \to 2$, $1 \to 3$
$2 \to 1$, $2 \to 2$,
$3 \to 2$, $3 \to 3$.



Определете PageRank на страниците.

Решение: Преходната матрица на тази марковска верига е

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

Стационареният вектор на разпределението

$$\pi = \begin{pmatrix}\pi_1 & \pi_2 & \pi_3\end{pmatrix}$$

удовлетворява уравненията

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

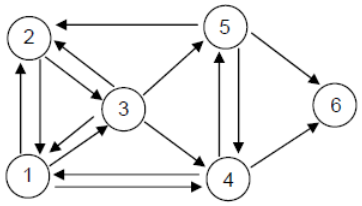
Като решим системата линейни уравнения, получаваме:

$$\begin{pmatrix}\pi_1 & \pi_2 & \pi_3\end{pmatrix} = \begin{pmatrix}\frac{3}{9} & \frac{4}{9} & \frac{2}{9}\end{pmatrix}.$$

Разбира се, реалната мрежа от уеб страници съдържа милиарди страници и съответно преходната матрица е с невероятна размерност, за да може така да просто да се изчисляват ранговете.

Да видим как Google използва стационарния вектор.

Пример 4. (Atherton [3]) Да разгледаме мрежата от 6 страници на Фиг. 5.



Фигура 5: Мрежа от 6 страници

Ще използваме тук наготово ранговете, определени чрез описания алгоритъм, а на читателя предлагаме да ги изчисли за упражнение. Стационарния вектор е

$$\pi = \begin{pmatrix} 0.2066 & 0.1770 & 0.1773 & 0.1770 & 0.1314 & 0.1309 \end{pmatrix}.$$

Нека си представим, че потребителят търси по ключови думи keyword1 и keyword2. Търсачката проверява в базата данни на Google с ключови думи, където за всяка ключова дума има списък от всички страници, в които тя се среща. Да предположим, че в базата данни се намира:

keyword1: page 2, page 5, page 6
keyword2: page 2, page 3
...

Резултата от търсенето ще са страниците $\{2, 3, 5, 6\}$. Следва сравнение на ранговете на тези страници, определени от PageRank, и подреждането им по важност. Ранговте на тези страници са $\pi_2 = 0.1770$, $\pi_3 = 0.1773$, $\pi_5 = 0.1314$ и $\pi_6 = 0.1309$. Следователно подреждането на страниците по ранг е:

$$page\ 3\ >\ page\ 2\ >\ page\ 5 >\ page\ 6.$$

В този ред са резултатите за потребителя. При ново търсене търсачката отново се обръща към в базата данни и извежда нов релевантен списък.

Справка с уебсайта на Google показва, че и досега PageRank е „сърцето на нашия софтуер". Трябва да се знае още, че PageRank не е единственият критерий, който Google използва за определяне на значимостта на уеб страниците. В действителност в момента алгоритъмът за подреждане е доста по-сложен и е търговска тайна.

Литература

[1] А. А. Марков (1906). Распространение закона больших чисел на величины, зависящие друг от друга. *Известия Физико-математического общества при Казанском университете*, 2-я серия, том 15, ст. 135-156.

[2] Е. Стоименова (2012). Случайно сърфиране в Интернет. *Математика и информатика*, година LV, кн. 3, 225-237.

[3] R. Atherton. *A look at Markov chains and their use in Google.* http://www.math.iastate.edu/thesisarchive/MSM/AthertonRMSMSS05.pdf (последен достъп на 11.08.2013)

[4] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems, 30(1-7):107-117, 1998.

[5] Ching, Ng, *Markov chains. Models, algorithms and applications*, 2006.

[6] Grinstead, Snell, *Introduction to probability*, 1997.

[7] L. Page, S. Brin, R. Motwani, T. Winograd (1998). The PageRank citation ranking: bringing order to the web. Technical report. Stanford Digital Library Technologies Project.

[8] Waner, Costenoble. *Finite mathematics*, 2011. (Chapter 7 Probability, Sec. 7.7 Markov systems.)