Вокруг Pagerank: марковские цепи, ранжирование веб-страниц и степенная экстраполяция

Сычев Сергей

Факультет Экономических Наук
НИУ ВШЭ

ИПС Методы линейной алгебры и анализа данных в экономике,
Ноябрь 2025

Содержание

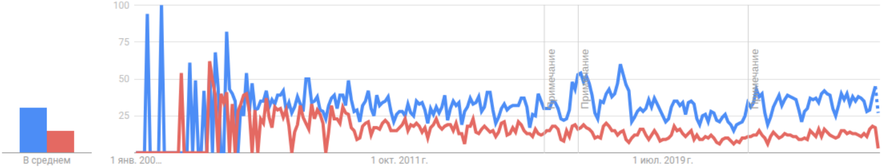
Содержание

Определение.

$A \in M_n(\mathbb{R})$ неприводимая $\iff$ $\nexists$ перестановочная матрица P:

$$P^{\mathsf{T}} A P = \begin{pmatrix} B & 0 \\ C & D \end{pmatrix}, \quad M_k(\mathbb{R}) \ni B, D \neq 0.$$

А зачем?



Использования Спектральной теорема и теоремы Перрона-Фробениуса

Теория из линейной алгебры

Теорема(Перрон-Фробениус).

Пусть $A \in M_n(\mathbb{R})$ — неприводимая неотрицательная матрица ($A \geq 0$).
Тогда:

(1) $\exists \lambda_{\max} > 0$ — собственное значение A, .

(2) $\lambda_{\max}$ имеет строго положительный собственный вектор $x > 0$.

Теория из линейной алгебры

Доказательство.

1. Пусть $A \in M_n(\mathbb{R})$ — неприводимая неотрицательная матрица.
2. Определим $S = \{x \geq 0, x \in \mathbb{R}^n, \|x\|_1 = 1\}$ — первый ортант и $S^+ = S \setminus \partial S$.
3. Введём отображение $f(x) = \frac{Ax}{\|Ax\|_1}$, которое переводит S в $S^+ \subset S$.
4. Неотрицательность и неприводимость гарантируют, что $Ax > 0$ для всех $x \in S$, следовательно $f(x) > 0$.
5. Так как S по построению - симплекс, то есть $co((e))$, где (e)-естественный базис $\mathbb{R}^n$ и f непрерывно, по теореме Какутани, существует неподвижная точка $x^* \in S$:
$$f(x^*) = x^* \implies Ax^* = \lambda_{\max} x^*, \quad \lambda_{\max} = \|Ax^*\|_1 > 0.$$
6. Таким образом, вектор x^* положителен (все координаты $x_i^* > 0$)

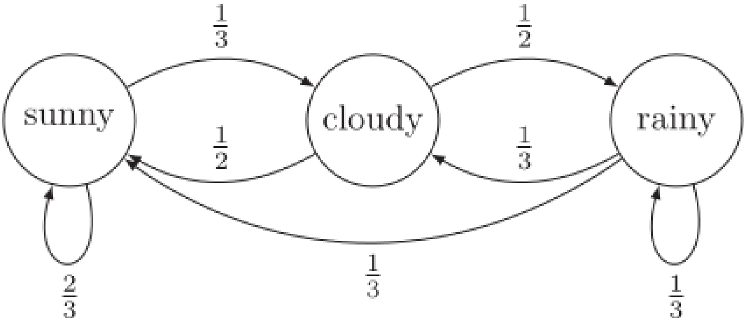
Замечание.

Более того,

- Алгебраическая(а значит и геометрическая) кратность$\lambda_{max} = 1$
- Если λ — любое другое собственное значение, то $|\lambda| \leq \lambda_{\text{max}}$.
- Если же $A > 0, \lambda$ — любое другое собственное значение, то $|\lambda| < \lambda_{\text{max}}$.

Содержание

Пример: А какая завтра погода?



Основные определения

Марковская цепь.

Марковской цепью называется последовательность случайных величин $\{\xi_t\}_{t=0}^{\infty}$ на (Ω, P) со значениями в не более чем счетном множестве X, называемом фазовым пространством, обладающая Марковским свойством:

$$\mathbb{P}(\xi_{t+1} = j \mid \xi_t = i, \xi_{t-1} = a_{t-1}, \ldots, \xi_0 = a_0) = \mathbb{P}(\xi_{t+1} = j \mid \xi_t = i).$$

Интуиция Марковского свойства.

Вероятность перехода в следующее состояние зависит только от текущего состояния, а не от всей истории процесса.

Переходные вероятности.

Для любых $i, j \in X$:

$$p_{ij} = \mathbb{P}(\xi_{t+1} = j \mid \xi_t = i).$$

Основные определения

Стохастическая матрица.

Квадратная матрица $P = [p_{ij}]$, где $p_{ij} \geq 0$ и $\sum_j p_{ij} = 1$ для всех i.

Однородная Марковская цепь.

Цепь называется однородной, если переходные вероятности p_{ij} не зависят от момента времени t.

Далее будем рассматривать исключительно однородные марковские цепи.

Распределение.

Вектор вероятностей

$$\pi_t = (\mathbb{P}(X_t = a_1), \ldots, \mathbb{P}(X_t = a_n)),$$

описывающий распределение состояний в момент времени t.

Основные определения

Стационарное распределение.

Вектор π называется стационарным, если

$$\pi P = \pi.$$

Непериодичность.

Цепь называется непериодичной, если для любого состояния i:

$$\text{НОД}\{t \geq 1 : (P^t)_{ii} > 0\} = 1.$$

Пример: Простейшее случайное блуждание

Случайное блуждание как Марковская цепь.

Пусть $\{\xi_t\}_{t \geq 0}, t \in \mathbb{Z}_+$ — независимые одинаково распределенные случайные величины с

$$\mathbb{P}(\xi_t = 1) = p, \quad \mathbb{P}(\xi_t = -1) = 1 - p.$$

Определим процесс

$$S_0 = 0, \quad S_t = \sum_{k=1}^{t} \xi_k.$$

Убедимся в выполнении Марковского свойства:

$$\mathbb{P}(S_{t+1} = i \mid S_t = j, \dots) = \mathbb{P}(\xi_{t+1} = i - j) = \mathbb{P}(S_{t+1} = i \mid S_t = j).$$

Эргодическая теорема

Теорема(Эргодическая).

Пусть $A \in M_n(\mathbb{R})$ — неприводимая неотрицательная стохастическая матрица. Тогда существует единственное стационарное распределение $\pi > 0$, такое что

$$P^\top \pi^\top = \pi^\top, \quad \sum_i \pi_i = 1.$$

Эргодическая теорема: доказательство

Доказательство.

1. Пусть λ — собственное значение P с собственным вектором $x^* \neq 0$, выберем координату $|{x^*}_k| = \max_i |{x^*}_i|$:
$$|\lambda||{x^*}_k| = |(P\mathbf{x}^*)_k| = \sum_j p_{kj}|{x^*}_j| \leq |{x^*}_k| \implies |\lambda| \leq 1.$$
2. Вектор $\overline{1} = (1, \ldots, 1)^\top$ является собственным вектором P, $\implies$ и $P^\top$:
$$P\overline{1} = \overline{1} \implies \lambda_{\max} = 1.$$
3. Неприводимость и неотрицательность $P \geq 0$, по теореме Перрона-Фробениуса, гарантируют существование строго положительного стационарного вектора $\pi > 0$, соответствующего собственному значению 1:
$$\pi P = \pi.$$

Содержание

Модель Интернета

Идея.

Будем считать Интернетом кортеж $(N, (O_i)_{i=1}^N, (I_i)_{i=1}^N)$,
где N - количество веб-страниц, O_i - количество исходящих ссылок со страницы i, I_i - количество входящих ссылок на страницу i.
Для упрощения рассуждений также положим, что с каждой страницы на другую может быть не более одной ссылки.
То есть Интернет - граф на N вершинах, который, вообще говоря,

1. непростой
2. несимметричный
3. неполный

Можно рассматривать посещенные страницы как однородную марковскую цепь в фазовом пространстве номеров страниц. В целом же в таких условиях удобно задать вероятность перехода со страницы j на страницу i как $\begin{cases} \frac{1}{O_j}, & \text{если существует ссылка } j \to i, \\ 0, & \text{иначе.} \end{cases}$

PageRank как задача поиска стационарного распределения

Ранги.

Отсюда естественно положить рангом i страницы

$$x_i = \sum_{k \in I_i} x_k p_{ik} = \sum_{k \in I_i} \frac{x_k}{O_k},$$

то есть ранги неизбежно зависят друг от друга. Перепишем в эквивалентной форме:

$$xP = x, (P)_{ij} = p_{ji} \qquad \sum_i x_i = 1, \quad x_i \geq 0.$$

Степенной метод

Итерации.

$$x^{(k+1)} = \frac{Ax^{(k)}}{\|Ax^{(k)}\|}, \quad x^{(0)} \neq 0.$$

Сходимость.

Если $|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n|$, то $x^{(k)} \to x_1$ — собственный вектор λ_1.

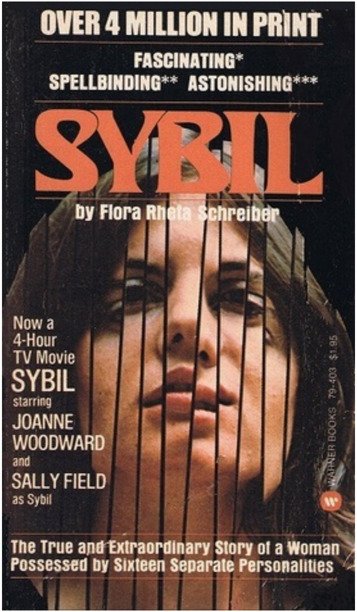
Скорость.

$$\|x^{(k)} - x_1\| = O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right).$$

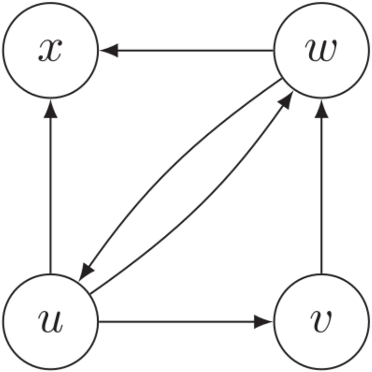
Тождество Рэлея.

$$\lambda^{(k)} = \frac{(x^{(k)})^\top Ax^{(k)}}{(x^{(k)})^\top x^{(k)}} \to \lambda_1.$$

Проблемы PageRank: Атака Сивиллы

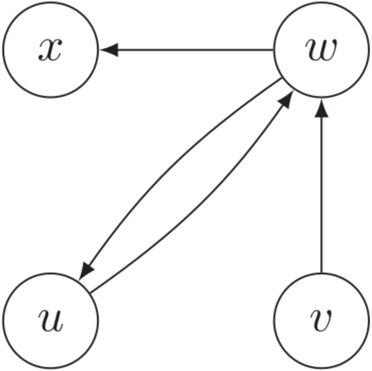


Проблемы PageRank: Атака Сивиллы



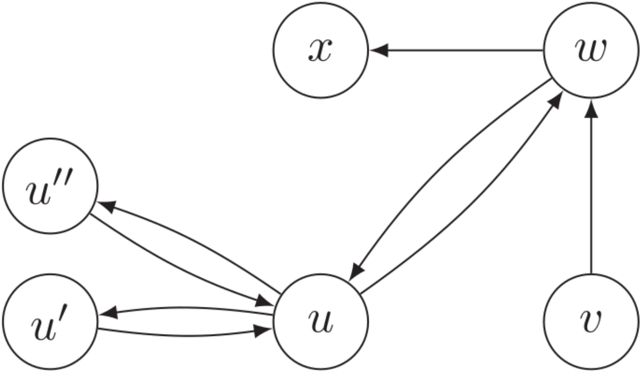
$$x_u \approx 0.21$$

Проблемы PageRank: Атака Сивиллы



$$x_u \approx 0.27$$

Проблемы PageRank: Атака Сивиллы

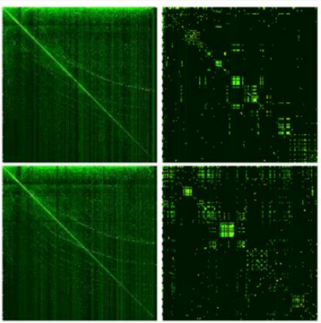


$x_u \approx 0.43$

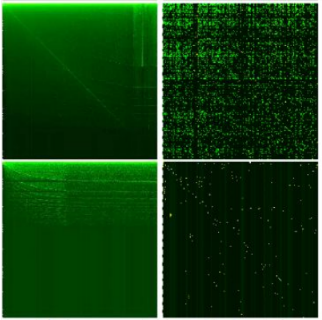
Проблемы PageRank: Сходимость степенного метода

Предложение.

Степенной метод сходится к некоторому вектору с наибольшему по модулю собственным значением, но значений, а значит и векторов, может быть несколько, если матрица, например, разреженная.



Cambridge 2006, University of Oxford 2006



Wikipedia English articles, PCN of Linux Kernel V2.6

А почему $\alpha = 0.85$?

Матрица Google.

$$G = \alpha W + (1-\alpha)[1]_n v^\top,$$

где $W = P^\top$ - матрица блуждающего, $v > 0$ - вектор персонализации, α - параметр демпфирования, обычно полагаем $\alpha = 0.85$. Такая матрица, по теореме Перрона-Фробениуса, гарантирует существование единственного собственного вектора с наибольшим по модулю собственным значением.

1. It just works
2. Иначе будут непропорциональные потери или в скорости сходимости, или в близости к стационарному распределению P
3. Вероятность пользователя продолжать блуждание:

Пусть $\eta : \Omega \to \mathbb{Z}_+$ - количество посещенных веб-страниц по цепочке.
По построению, $\eta \sim Geom(1-\alpha)$.
Знаем, что $E_\eta = \frac{1}{1-\alpha}$.
По наблюдениям, $E_\eta = \overline{6,7} \in \mathbb{R}$, что достигается как раз при $\alpha \approx 0.85$.

Содержание

А нельзя ли как-то ускорить?

Идея: использовать накопленные итерации $x^{(k-2)}, \dots, x^{(1)}$ Будем искать $x^{(k-1)} = u_1 + a_2 u_2$, где u_1, u_2 сопоставлены собственные значения 1 и $c = 0.85$ соответственно. Отсюда

$$\frac{u_1 = x^{(k)} - cx^{(k-1)}}{1 - c}$$

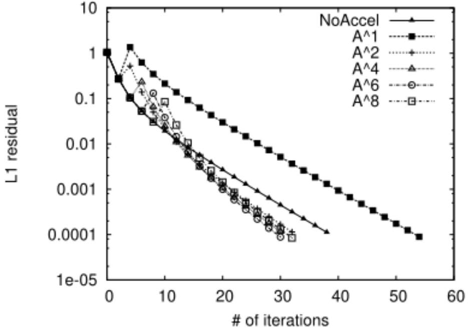
Но такая экстраполяция будет сходиться медленнее, учтено лишь собственное значение с собственным вектором c, хотя могут быть и $-c, ci, -ci, \dots$!

Поэтому рассмотрим $x^{(k-1)} = u_1 + a_2 u2 + a_3 u_3$, где u_1, u_2 сопоставлены собственные значения $1, c = 0.85$ и $- = -0.85$ соответственно. Получаем

$$\frac{u_1 = x^{(k)} - c^2 x^{(k-2)}}{(1 - c)^2},$$

что дает ускорение работы алгоритма на $\approx 18\%$

Ускорение степенного метода: степенная экстраполяция



(a) c=0.85

Type	speedup
$d = 1$	-28%
$d = 2$	18%
$d = 4$	25.8%
$d = 6$	30%
$d = 8$	21.8%
Quadratic	20.8%

Содержание

Библиография

- G. H. Golub, C. F. Van Loan, *Matrix Computations*, 4th edition, 2013.
- Stanford NLP, *Power Extrapolation for Markov Chains*, `https://nlp.stanford.edu/~manning/papers/PowerExtrapolation.pdf`
- S. Brin, L. Page, *The Anatomy of a Large-Scale Hypertextual Web Search Engine*, 1998.
- Ф. Р. Гантмахер. *Теория матриц*. — Москва: Наука, 1966.
- Larry Page. *PageRank: Bringing Order to the Web*. — Stanford University, 1996.
- Taher Haveliwala, Sepandar Kamvar, Dan Klein, Chris Manning, Gene Golub. *Computing PageRank using Power Extrapolation*. — Stanford University, 2003.
- Александров Артём, Удальцов Валентин. *Принцип ранжирования интернет-страниц поисковыми системами*. — 2012.
- Yen Do, Hoi Nguyen, Van Vu. *Real Roots of Random Polynomials: Expectation and Repulsion*. — Annals of Probability, 2014.
- Paolo Boldi, Massimo Santini, Sebastiano Vigna. *PageRank as a Function of the Damping Factor*. — WWW Conference, 2005.