Итоги семинара 24.11.2020

На семинаре мы продолжили изучение цепей Маркова и познакомились с двумя новыми их свойствами: неразложимостью и апериодичностью.

Цепь Маркова (ЦМ) – это последовательность случайных чисел $\{X_n\}_{n\geqslant 0}$ с конечным (или счетным) числом исходов, (то есть в каждый момент времени система может принимать одно из состояний $E = \{0, 1, 2, \ldots\}$, обладающая марковским свойством:

$$\mathbb{P}\left(X_{n+1} = j \mid X_0 = i_0, \ldots, X_{n-1} = i_{n-1}, X_n = i\right) = \mathbb{P}\left(X_{n+1} = j \mid X_n = i\right)$$

для любого n и любых состояний $i, j, i_0, \ldots, i_{n-1} \in E$.

ЦМ называется однородной (по времени), если вероятности перехода из i-го состояния в j-е не зависят от момента времени n, в который мы наблюдаем процесс. Обозначим эти вероятности через p_{ij}:

$$p_{ij} = \mathbb{P}\left(X_1 = j \mid X_0 = i\right) = \mathbb{P}\left(X_{n+1} = j \mid X_n = i\right).$$

Матрица $\mathbf{P} = (p_{ij})_{i,j\geqslant 1}$ называется матрицей переходных вероятностей.

Обозначим $\pi_j^0 = \mathbb{P}\left(X_0 = j\right)$. Вектор $\pi^0 = (\pi_1^0, \pi_2^0, \ldots, \pi_N^0)$ называется *распределением цепи в начальный момент времени*. Аналогично определяются вектора $\pi^1, \pi^2, \ldots$:

$$\begin{aligned} \pi_j^n &= \mathbb{P}\left(X_n = j\right) \\ \pi^n &= (\pi_1^n, \pi_2^n, \ldots, \pi_N^n). \end{aligned}$$

Можно показать, что

$$\begin{aligned} \pi^1 &= \pi^0 \mathbf{P} \\ \pi^n &= \pi^{n-1}\mathbf{P} = \cdots = \pi^0 \mathbf{P}^n. \end{aligned}$$

Распределение называется *стационарным*, если $\pi = \pi\mathbf{P}$.

Цепь Маркова называется *неразложимой*, если из каждого состояния можно перейти в любое другое с положительной вероятностью (за некоторое количество шагов).

Цепь Маркова называется *апериодичной*, если апериодичио любое ее состояние j : не существует такого числа $d > 1$, что любой путь из j в j имеет длину, кратную d. Для проверки апериодичности неразложимой цепи Маркова достаточно убедиться в апериодичности одного любого ее состояния.

Теорема 1 *Если конечная однородная ЦМ неразложима и апериодична, то она имеет единственное стационарное распределение π и при $n \to \infty$*

$$\pi_j^n \to \pi_j, \quad j = 1, \ldots, N$$

PageRank

Алгоритм PageRank решает задачу ранжирования поисковой выдачи. Основная идея алгоритма – наиболее релевантными должны быть страницы, имеющие большую «важность», то есть на них часто ссылаются авторитетные источники, которые тоже сами часто посещаются.

Чтобы посчитать PageRank для страниц в сети, мы задаём цепь Маркова: страницы — это возможные состояния, переходные вероятности задаются ссылками со страницы на страницу (взвешенными таким образом, что на каждой странице все связанные страницы имеют одинаковую вероятность выбора). Модифицировав такую ЦМ некоторым образом, мы получаем неразложимую и апериодичную цепь, у которой можно искать стационарное распределение. Алгоритм будет ранжировать страницы по убыванию вероятности перехода на эти страницы в стационарном распределении.

Такая версия алгоритма слишком проста и поддается «взлому». Метод взлома, который мы рассмотрели, называется sybil attack. Было показано, как владелец любой страницы может сделать ее PR сколь угодно большим с помощью «фермы» из страниц-клонов.

Что бы еще почитать?

Общая теория цепей Маркова изложена в классическом учебнике [1] и [2]. Собственно, статья Сергея Брина и Ларри Пейджа, в которой впервые были описаны поисковая система Google и алгоритм PageRank [3]. Ее перевод на русский – [4].

Для продвинутых: обзор стратегий типа Sybil обмана алгоритма PageRank [5]. Еще можно просто погуглить PageRank: написано куча статей, где разбираются алгоритмы обмана старых версий, предлагаются улучшения и обсуждаются вычислительные сложности.

Литература

[1] Ширяев, Альберт Николаевич. Вероятность-1. МЦНМО, 2007.

[2] Чжун, Кай-лай. Однородные цепи Маркова. Мир, 1964.

[3] Brin, Sergey, and Lawrence Page. "The anatomy of a large-scale hypertextual web search engine."(1998).

[4] `https://web.archive.org/web/20130625003347/http://wseob.ru/seo/searchengine-anatomy`

[5] Cheng, Alice, and Eric Friedman. "Manipulability of PageRank under sybil strategies."(2006).