ПРИМЕНЕНИЕ СТАЦИОНАРНЫХ ЦЕПЕЙ МАРКОВА ДЛЯ ОПТИМИЗАЦИИ ПОИСКОВЫХ СИСТЕМ

Башкевич А. Ю., Жук Е. Е.
Факультет прикладной математики и информатики, кафедра математического моделирования и анализа данных, Белорусский государственный университет
Минск, Республика Беларусь
E-mail: bashkevichay@gmail.com, zhukee@mail.ru

В данной работе предлагается использовать модель стационарной односвязной цепи Маркова для подсчета значения PageRank при оптимизации поисковой системы Google. PageRank является числом, отображающим важность страницы. При поиске и ранжировании документов поисковая система Google основывается на содержании страницы, ключевых словах в заголовке и описании, после этого на положение страницы влияет ее PageRank.

Введение

Если рассматривать поведение абстрактного пользователя сети, который выбирает ссылки случайным образом и при этом в любой момент может закрыть браузер и прекратить просмотр, то возникает вопрос, с какой вероятностью случайный пользователь попадет на ту или иную страницу и от чего эта вероятность зависит. Именно для расчета вероятности посещения страницы был разработан алгоритм PageRank.

I. Алгоритм ссылочного ранжирования

Для расчета PageRank [1] используется формула (1), где $PR(T_i)$ – PageRank страницы T_i, $T_1, ..., T_{i-1}, T_{i+1}, ..., T_n$ – страницы, ссылающиеся на страницу T_i , $C(T_i)$ – количество ссылок страницы T_i , d – коэффициент затухания, находится в пределах от 0 до 1 (обычно равен 0,85), он введен для учета вероятности того, что пользователь закроет страницу. Также на PageRank наложено ограничение [1]:

$$\sum_{i=1}^{N} PR(T_i) = N, \qquad (2)$$

где N – количество страниц в сети. Данное условие следует из того, что сумма всех вероятностей не может превышать единицу, т.е. вероятность пребывания на данной странице равна отношению значения PageRank к числу всех страниц.

Рассмотрим расчет PageRank для простейшей сети, состоящей из четырех страниц.

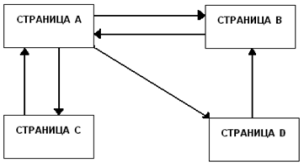

Рис. 1 – Пример страниц в сети

Теперь посчитаем значения PageRank на первом шаге, подставив $PR(B)$ и $PR(C)$ равные 1 в формулу (1). В результате мы получим новые значения PageRank для всех страниц. Теперь посчитаем значения PageRank на втором этапе. Здесь подставим $PR(B)$ и $PR(C)$, полученные на предыдущем шаге. Таким образом, мы получим значения PageRank на втором шаге, которые будут использоваться при расчете значений PageRank на третьем шаге.

Смысл заключается в том, что нам придется проделать эту операцию снова и снова, используя каждый раз значения PageRank рассчитанные на предыдущем этапе. И в результате после какого-то шага n на всех последующих шагах, начиная с $n + 1$, значения PageRank будут неизменными. Для нашего примера достаточно проделать 10 шагов, чтобы значения PageRank выглядели следующим образом: $PR(A) = 1,637; PR(B) = 1,136; PR(C) = 0,614; PR(D) = 0,614$. Сумма всех PR равна 4, т.е. условие (2) выполняется.

Также PageRank можно рассчитать матричным методом. Для этого составляется матрица следующего вида:

$$\begin{array}{c} \text{A} \quad \text{B} \quad \text{C} \quad \text{D} \\ \begin{pmatrix} 0 & 1/3 & 1/3 & 1/3 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

Данная матрица соответствует нашей простейшей сети (см. рис. 1), т. е. страница A ссылается на B, C, D. Страница B ссылается на A. Страница C ссылается на A и D ссылается на B. При этом значения каждой строки делятся на количество ссылок данной страницы.

Данную матрицу необходимо умножить на значение PR с предыдущего шага, полученный вектор умножить на единичный вектор, умноженный на d, и прибавить к результату единичный вектор, умноженный на $(1 - d)$.

После расчета мы видим, что страница A имеет самый высокий PR в нашей сети, страница B – более низкий. Поэтому если все четыре страницы будут по содержанию соответствовать какому-то запросу поиска, то после учета значений PR страница A окажется на первом месте, B – на втором, а C и D – на третьем.

II. Цепи Маркова

Определение 1. Временной ряд x_1, ..., x_T, x_{T+1},...ϵS, где S=$\{1, ..., L\}$, называется однородной цепью Маркова [2] с пространством состояний S, образованной из $L \geq 2$ состояний, если выполняется так называемое марковское свойство (3) $\forall d_{t+1}, ...\epsilon S, \ \forall t = 1, ..., T, ...$

Две вероятностные характеристики цепи Маркова [2 3]:

1. *начальное распределение вероятностей:*

$$\pi_i^{(1)} = P(x_1 = i), \ i\epsilon S; \ \sum_{i\epsilon S} \pi_i^{(1)} = 1; \quad (4)$$

2. *матрицу вероятностей одношаговых переходов (5).*

Цепь Маркова характеризуется своим текущим распределением вероятностей (в момент времени t):

$$\pi_i^{(t)} = P\{x_t = i\}, \ i\epsilon S, \ t = 1, 2, ..., T, ... \quad (6)$$

Теорема 1. В условиях модели (3) – (5) текущее распределение вероятностей (6) может быть найдено из соотношения [2]:

$$\pi^{(t)} = (P')^{t-1}\pi^{(1)} = (P^{t-1})\pi^{(1)}, \quad (7)$$

где $\pi^{(t)} = (\pi_1^{(t)}, ..., \pi_L^{(t)}), \ t = 1, 2, ..., T, ...$ а $\pi^{(1)} = \pi^{(t)}|_{t=1}$ – начальное распределение вероятностей.

Определение 2. Распределение вероятностей $\pi = (\pi_1, ..., \pi_L)$ называется стационарным распределением для однородной цепи Маркова (3) – (5), если выполняется [2]:

$$\begin{cases} P'\pi = \pi; \\ \sum_{i\epsilon S} \pi_i = 1. \end{cases} \quad (8)$$

Теорема 2. Пусть для однородной цепи Маркова (3) – (5) начальное распределение вероятностей $\pi^{(1)}$ совпадает со стационарным из (8): $\pi^{(1)} = \pi$, тогда данная цепь Маркова является стационарной, и текущее распределение вероятностей в любой момент времени совпадает со стационарным [2]:

$$\pi^{(1)} = \pi, \ t = 1, 2, ..., T, ... \quad (9)$$

III. Расчет показателя PageRank с помощью цепей Маркова

Следуя из определения цепи Маркова можно сказать, что описанный выше процесс поведения абстрактного пользователя сети является марковским. Начальное состояние системы описывается вектором начальных вероятностей цепи Маркова.

Если $\pi^{(0)}$ это вектор начальных вероятностей, то $\pi^{(t)}$ – это вероятностный вектор на шаге t. Значение $\pi^{(t+1)}$ расчитываем по формуле (7). Если цепь Маркова стационарная, то при t стремящемся к бесконечности, вероятности состояний стремятся к определенным предельным значениям, которые удовлетворяют следующему соотношению:

$$\pi^{(t)} \to \pi. \quad (10)$$

Из (10) очевидно, что π является единственным и не зависит от вектора начальных вероятностей $\pi^{(0)}$ и определяется только матрицей вероятностей переходов (5). Матрица вероятностей одношаговых переходов будет выглядеть также как и при расчете PageRank матричным способом.

Для нашей системы вектор начальных вероятностей выглядит следующим образом (поскольку в нашей сети четыре страницы и пользователь может оказаться на любой из них с одинаковой вероятностью): $\pi^{(0)} = (1/4, 1/4, 1/4, 1/4)$.

Для расчета вектора $\pi^{(1)}$ применим формулу (7). Для вектора $\pi^{(2)}$ формула (7) имеет следующий вид: $\pi^{(2)} = P'\pi^{(1)}$. Далее аналогично рассчитывая значения $\pi^{(t)}$ по формуле (7) мы найдем t, для которого выполняется равенство (10). Т.е. мы найдем значение вектора π = (0,429; 0,286; 0,143; 0,143).

Таким образом, после достаточно большого количества переходов со страницы на страницу уже не имеет значения с какой страницы пользователь начал просмотр, поскольку в результате он окажется на странице A с вероятностью 0,429, на странице B с вероятностью 0,286, на страницах C и D с вероятностями 0,143.

Теперь с полученным вектором π произведем вычисления по формуле (1): 0,85*4* π + (1-0,85). Получается что, мы получили значения практически равные значениям PR, рассчитанным в первой части статьи.

IV. Список литературы

1. Студия дизайна МЕДИАКРАФТ [Электронный ресурс] / Медиакрафт. – Москва, 2007. – Режим доступа: http://www.mediacraft.ru/sections/articles/1/1.html. – Дата доступа: 08.09.2013.
2. Харин, Ю. С. Математическая и прикладная статистика / Ю. С. Харин, Е. Е. Жук – Мн.: БГУ, 2005. – 279 с.
3. Кемени, Дж. Дж. Конечные цепи Маркова / Дж. Дж. Кемени, Дж. Л. Снелл –М.: Наука, 1970. – 272 с.

$$PR(T_i) = (1-d) + d(\frac{PR(T_1)}{C(T_1)} + ... + \frac{PR(T_{i-1})}{C(T_{i-1})} + \frac{PR(T_{i+1})}{C(T_{i+1})} + ... + \frac{PR(T_N)}{C(T_N)}) \quad (1)$$

$$P\{x_{t+1} = d_{t+1}|x_t = d_t, x_{t-1} = d_{t-1}, ..., x_1 = d_1\} = P\{x_{t+1} = d_{t+1}|x_t = d_t\} = p_{d_t.d_{t+1}} \quad (3)$$

$$P = (p_{ij})_{i,j\epsilon S}, \ p_{ij} = P\{x_{t+1} = j|x_t = i\}, \ \sum_{j\epsilon S} \pi_{ij} = 1, \ i\epsilon S \quad (5)$$