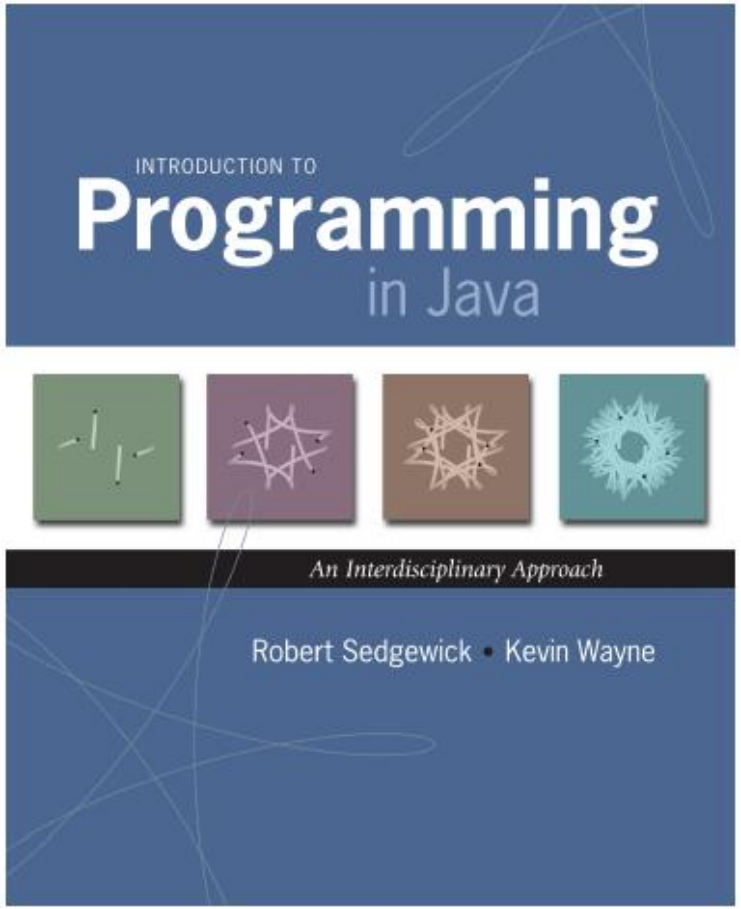
عناصر برنامه‌نویسی: کاوشگر تصادفی وب

سید ناصر رضوی n.razavi@tabrizu.ac.ir

۱۳۹۵

۱-۶ کاوشگر وب

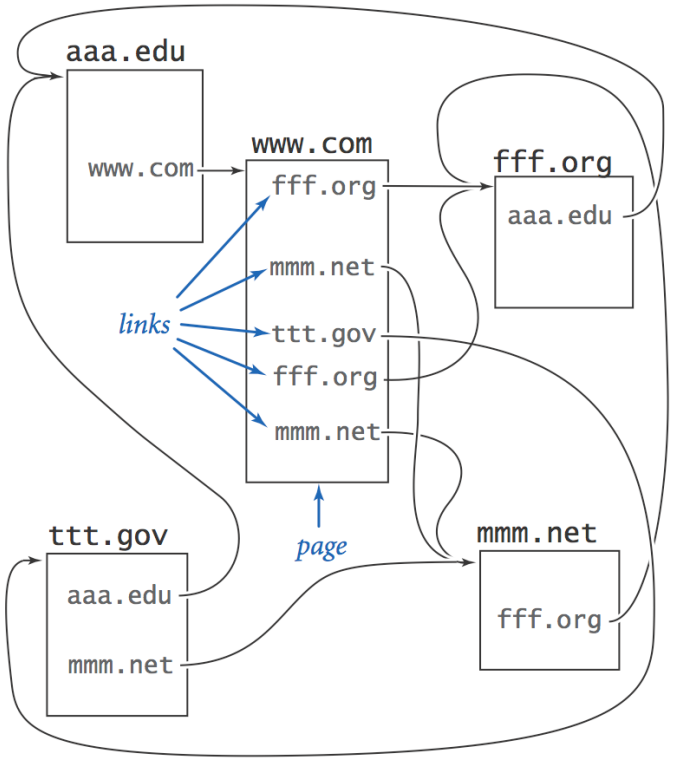

INTRODUCTION TO
Programming
in Java
An Interdisciplinary Approach
Robert Sedgewick • Kevin Wayne

جستجوی وب

- مرتبط بودن. آیا سند مفروض شبیه واژه‌های پرس و جو است؟
- اهمیت. آیا سند مفروض برای کاربرهای گوناگون مفید است؟

الگوریتم PageRank

- الگوریتم PageRank گوگل. [سرجی برین و لری پیج، ۱۹۹۸]
  - اندازه‌گیری محبوبیت صفحات بر اساس ساختار پیوندی وب.
  - یک انقلاب در زمینه‌ی دسترسی به اطلاعات.

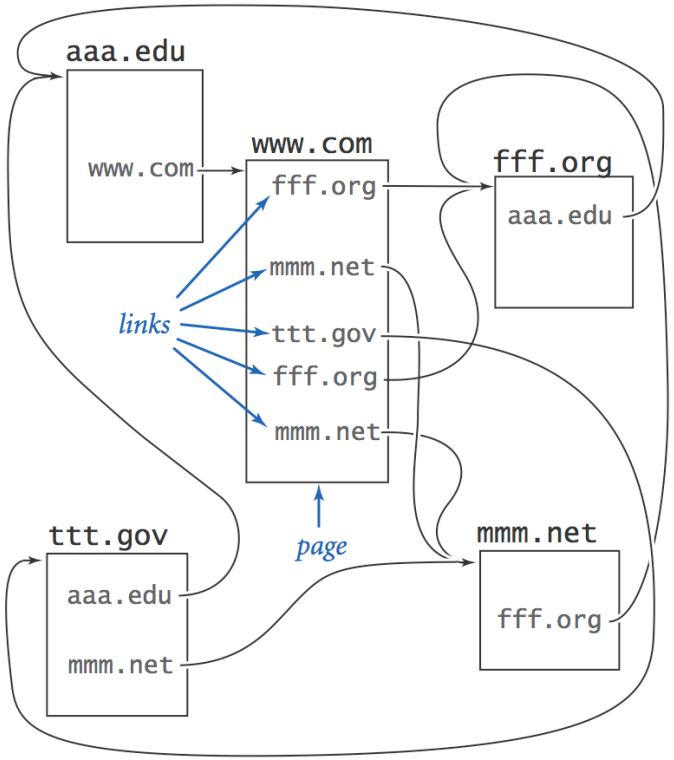


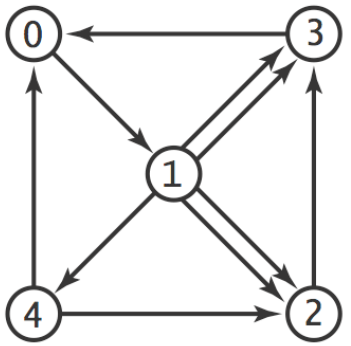
قاعده ۱۰-۹۰

- مدل. انتخاب صفحه بعدی به وسیله کاوشگر وب:
  - ۹۰ درصد مواقع کاوشگر بر روی یک پیوند تصادفی کلیک می‌کند.
  - ۱۰ درصد مواقع کاوشگر به یک صفحه کاملاً تصادفی می‌رود.
- هشدارها. یک مدل خام اما مفید برای کاوش صفحات وب.
  - هیچ کس پیوندها را با احتمال تصادفی انتخاب نمی‌کند.
  - قاعده ۱۰-۹۰ تنها یک حدس است.
  - این مدل دکمه بازگشت به صفحه قبل و همچنین صفحات نشانه‌گذاری شده را در نظر نمی‌گیرد.
  - تنها می‌تواند از عهده یک نمونه‌ی بسیار کوچک از وب برآید.
  - ...

قالب ورودی برای گراف وب

- قالب ورودی.
  - N صفحه با شماره‌های 0 تا N - 1
  - نمایش هر پیوند به وسیله یک زوج از اعداد صحیح





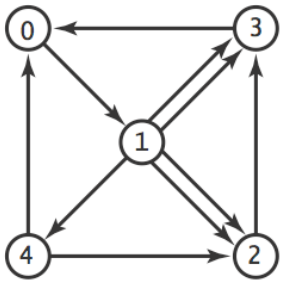

```
% more tiny.txt
5          ← N
0 1
1 2   1 2
1 3   1 3   1 4
2 3
3 0
4 0   4 2
```

links

ماتریس انتقال

- ماتریس انتقال. `p[i][j]` = احتمال رفتن از صفحه `i` به صفحه `j`

input graph




```
5
0 1
1 2   1 2
1 3   1 3   1 4
2 3
3 0
4 0   4 2
```

link counts

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 2 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \end{bmatrix}$$

degrees

$$\begin{bmatrix} 1 \\ 5 \\ 1 \\ 1 \\ 2 \end{bmatrix}$$

leap probabilities

$$\begin{bmatrix} .02 & .02 & .02 & .02 & .02 \\ .02 & .02 & .02 & .02 & .02 \\ .02 & .02 & .02 & .02 & .02 \\ .02 & .02 & .02 & .02 & .02 \\ .02 & .02 & .02 & .02 & .02 \end{bmatrix} + \begin{bmatrix} 0 & .90 & 0 & 0 & 0 \\ 0 & 0 & .36 & .36 & .18 \\ 0 & 0 & 0 & .90 & 0 \\ .90 & 0 & 0 & 0 & 0 \\ .45 & 0 & .45 & 0 & 0 \end{bmatrix}$$

link probabilities

$$= \begin{bmatrix} .02 & .92 & .02 & .02 & .02 \\ .02 & .02 & .38 & .38 & .20 \\ .02 & .02 & .02 & .92 & .02 \\ .92 & .02 & .02 & .02 & .02 \\ .47 & .02 & .47 & .02 & .02 \end{bmatrix}$$

transition matrix

احتمال رفتن از صفحه ۱ به صفحه ۲

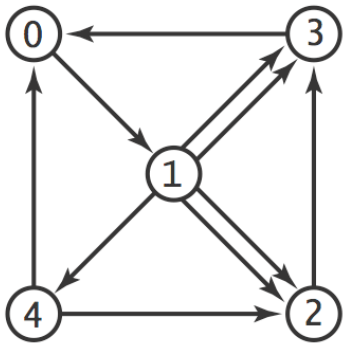
محاسبه ماتریس انتقال از روی گراف وب


```
public class Transition {
   public static void main(String[] args) {
      int N = StdIn.readInt();                          // # number of pages
      int[][] counts = new int[N][N];                   // # links from page i to j
      int[] outDegree = new int[N];                     // # links from page

      // accumulate link counts
      while (!StdIn.isEmpty()) {
         int i = StdIn.readInt();
         int j = StdIn.readInt();
         outDegree[i]++;
         counts[i][j]++;
      }

      // print transition matrix
      StdOut.println(N + " " + N);
      for (int i = 0; i < N; i++) {
         for (int j = 0; j < N; j++) {
            double p = .90 * counts[i][j] / outDegree[i] + .10/N;
            StdOut.printf("%7.5f ", p);
         }
         StdOut.println();
      }
   }
}
```

محاسبه ماتریس انتقال از روی گراف وب




```
% more tiny.txt
5 ← N
0 1
1 2   1 2
1 3   1 3   1 4
2 3
3 0
4 0   4 2
```

links

```
% java Transition < tiny.txt
5 5
 0.02000 0.92000 0.02000 0.02000 0.02000
 0.02000 0.02000 0.38000 0.38000 0.20000
 0.02000 0.02000 0.02000 0.92000 0.02000
 0.92000 0.02000 0.02000 0.02000 0.02000
 0.47000 0.02000 0.47000 0.02000 0.02000
```

شبیه‌سازی مونت کارلو

- شبیه‌سازی مونت کارلو.
  - کاوشگر از صفحه صفر شروع می‌کند.
  - به صورت تکراری و بر اساس ماتریس انتقال صفحه بعدی را انتخاب می‌کند.
  - سپس محاسبه می‌کند از هر صفحه به چه میزان بازدید شده است.

چگونه؟ اسلاید بعد را ببینید

$$\begin{bmatrix} .02 & .92 & .02 & .02 & .02 \\ .02 & .02 & .38 & .38 & .20 \\ .02 & .02 & .02 & 92 & .02 \\ .92 & .02 & .02 & .02 & .02 \\ .47 & .02 & .47 & .02 & .02 \end{bmatrix}$$

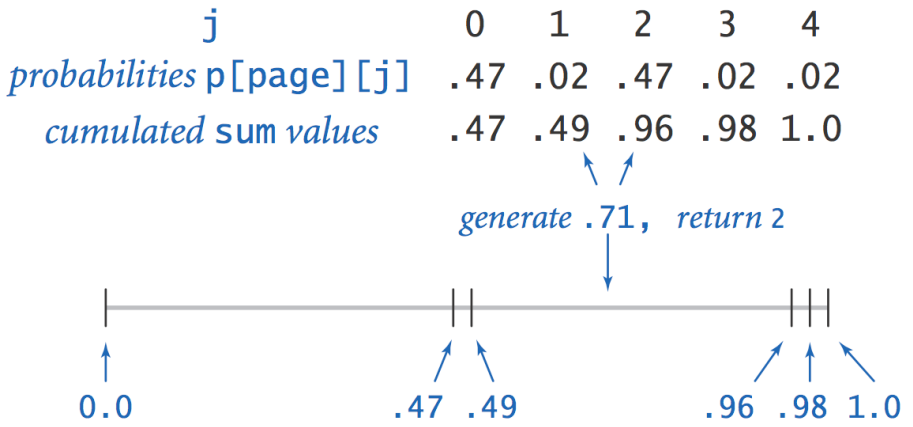
ماتریس انتقال

کاوشگر تصادفی

- حرکت تصادفی. فرض کنید کاوشگر در صفحه‌ی page است؟ صفحه‌ی بعدی j چگونه انتخاب می‌شود؟
  - سطر page از ماتریس انتقال شامل احتمالات مربوط به این صفحه است.
  - احتمال‌های تجمعی برای این صفحه محاسبه می‌شود.
  - یک عدد تصادفی مانند r در بازه‌ی 0.0 و 1.0 تولید می‌شود.
  - صفحه بعدی j بر اساس بازه‌ای که r در آن قرار گرفته انتخاب می‌شود.

	.02	.92	.02	.02	.02
	.02	.02	.38	.38	.20
	.02	.02	.02	.92	.02
	.92	.02	.02	.02	.02
page	.47	.02	.47	.02	.02

ماتریس انتقال



کاوشگر تصادفی

- حرکت تصادفی. فرض کنید کاوشگر در صفحه‌ی page است؟ صفحه بعدی j چگونه انتخاب می‌شود؟
  - سطر page از ماتریس انتقال شامل احتمالات مربوط به این صفحه است.
  - احتمال‌های تجمعی برای این صفحه محاسبه می‌شود.
  - یک عدد تصادفی مانند r در بازه 0.0 و 1.0 تولید می‌شود.
  - صفحه بعدی j بر اساس بازه‌ای که r در آن قرار گرفته انتخاب می‌شود.

```
// make one random move
double r = Math.random();
double sum = 0.0;
for (int j = 0; j < N; j++) {
   // find interval containing r
   sum += p[page][j];
   if (r < sum) { page = j; break; }
}
```

کاوشگر وب: شبیه‌سازی مونت کارلو


```
public class RandomSurfer {
   public static void main(String[] args) {
      int T = Integer.parseInt(args[0]);          // number of moves
      int N = StdIn.readInt();                    // number of pages
      int page = 0;                               // current page
      double[][] p = new int[N][N];               // transition matrix

      // read in transition matrix
      ...

      // simulate random surfer and count page frequencies
      int[] freq = new int[N];
      for (int t = 0; t < T; t++) {
         // make one random move
         freq[page]++;
      }

      // print page ranks
      for (int i = 0; i < N; i++) {
         StdOut.printf("%8.5f", (double) freq[i] / T);
      }
      StdOut.println();
   }
}
```

اسلاید قبل

رتبه‌ی صفحه

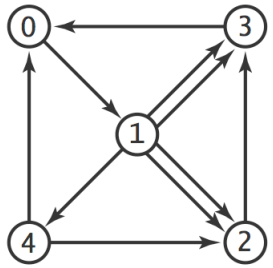
همگرایی: کمی ریاضیات

- همگرایی. برای مدل کاوشگر تصادفی، مدت زمانی که کاوشگر بر روی هر صفحه صرف می‌کند به یک توزیع منحصر بفرد همگرا می‌شود که مستقل از صفحه شروع است.

رتبه‌ی صفحه

توزیع ایستای زنجیره‌ی مارکوف

بردارهای ویژه‌ی اصلی ماتریس انتقال

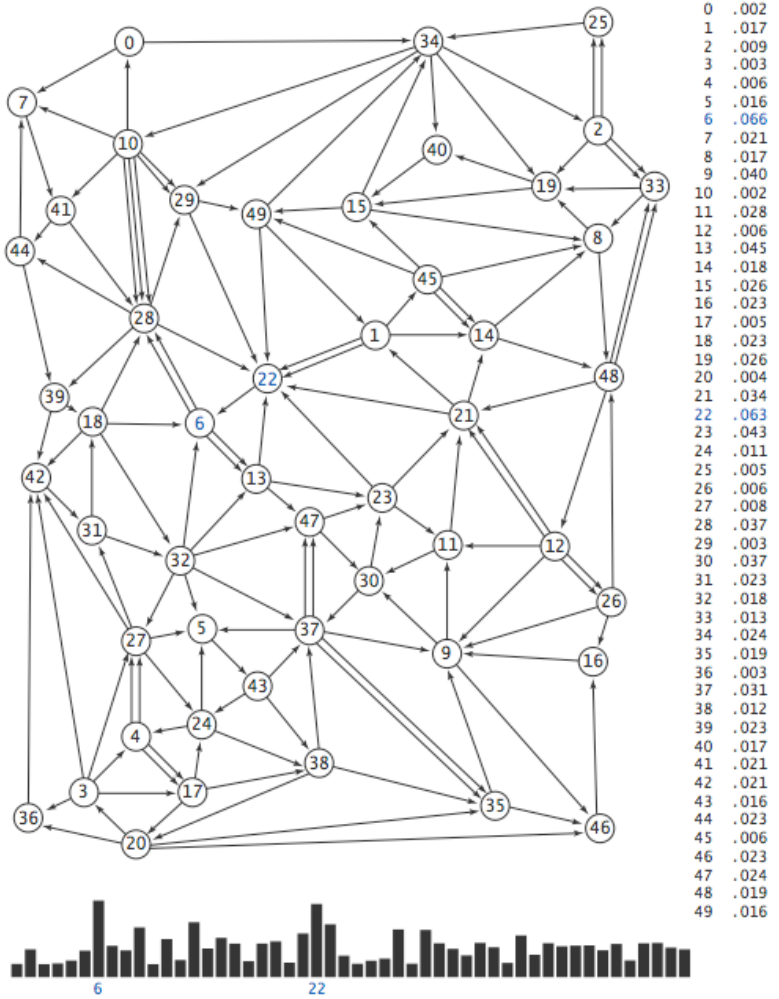


0 .27
1 .27
3 .25
2 .15
4 .07

$$\left[\frac{428{,}671}{1{,}570{,}055} \quad \frac{417{,}205}{1{,}570{,}055} \quad \frac{229{,}519}{1{,}570{,}055} \quad \frac{388{,}162}{1{,}570{,}055} \quad \frac{106{,}498}{1{,}570{,}055}\right]$$

یک گراف بزرگ‌تر


```
% more medium.txt
50
 0  7  0 34
 1 14  1 22  1 22  1 45
 2 19  2 25  2 33
 3  4  3 17  3 27  3 36  3 42
 4 17  4 17  4 27  4 27
 5 43
 6 13  6 13  6 28
 7 41
 8 19  8 48
 9 11  9 30  9 46
 ...
```



Page ranks with histogram for a larger example