Марковские цепи в задаче ранжирования страниц в Интернете

Т.В. Жуковская, доцент

ТГТУ, кафедра высшей математики

План доклада

- Марковские цепи

План доклада

- Марковские цепи
- Задача ранжирования

План доклада

- Марковские цепи
- Задача ранжирования
- Алгоритм PageRank

План доклада

- Марковские цепи
- Задача ранжирования
- Алгоритм PageRank
  - Модель случайного блуждания

План доклада

- Марковские цепи
- Задача ранжирования
- Алгоритм PageRank
  - Модель случайного блуждания
  - Алгоритм вычисления

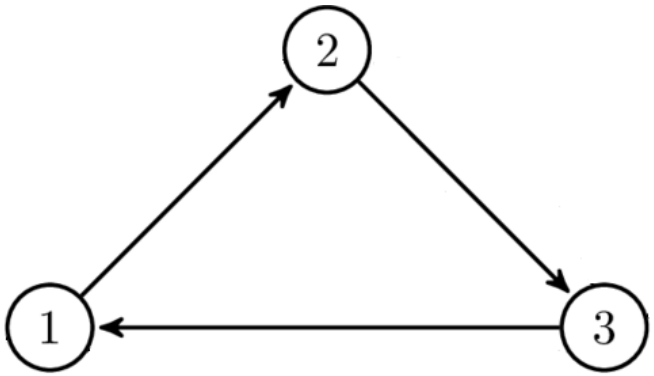
План доклада

- Марковские цепи
- Задача ранжирования
- Алгоритм PageRank
  - Модель случайного блуждания
  - Алгоритм вычисления
  - Применение к задаче ранжирования

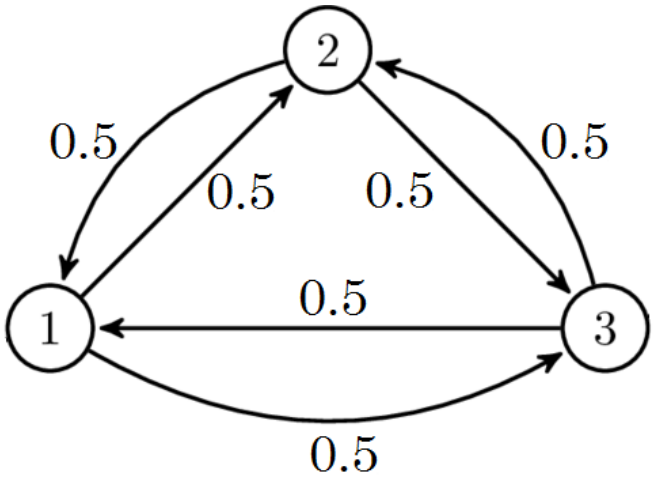
План доклада

- Марковские цепи
- Задача ранжирования
- Алгоритм PageRank
  - Модель случайного блуждания
  - Алгоритм вычисления
  - Применение к задаче ранжирования
- Другие модели, основанные на марковских цепях

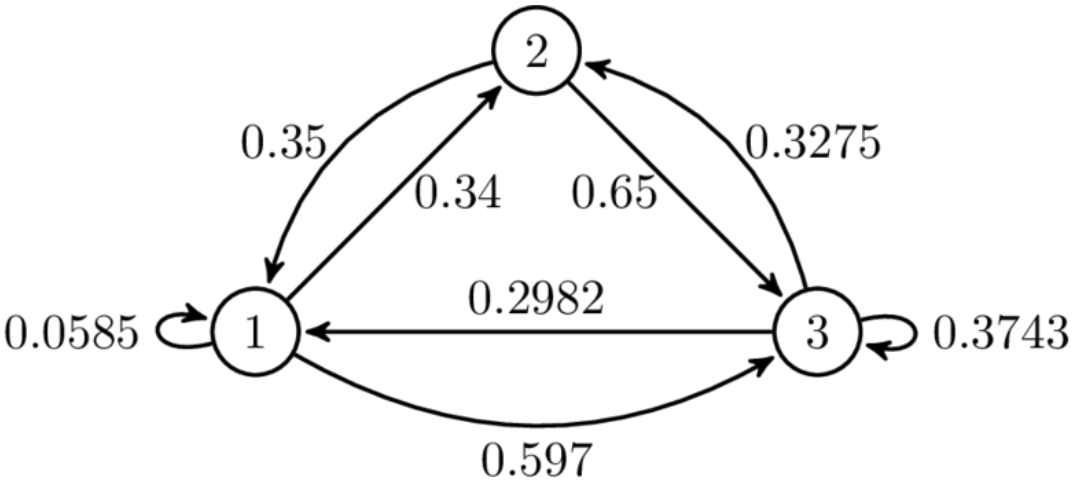
Марковские цепи: идеология



Марковские цепи: идеология



Марковские цепи: идеология



Цепь Маркова с дискретным временем — случайный процесс $\{X_n\}_{n \geq 0}$:

$$\mathsf{P}(X_{n+1} = i_{n+1} | X_n = i_n, \ldots, X_0 = i_0) =$$

$$\mathsf{P}(X_{n+1} = i_{n+1} | X_n = i_n).$$

Однородная: $\mathsf{P}(X_{n+1} = j | X_n = i) = \mathsf{P}(X_1 = j | X_0 = i)$.

Цепь Маркова с дискретным временем — случайный процесс $\{X_n\}_{n\geq 0}$:

$$\mathsf{P}(X_{n+1} = i_{n+1} | X_n = i_n, \ldots, X_0 = i_0) =$$

$$\mathsf{P}(X_{n+1} = i_{n+1} | X_n = i_n).$$

Однородная: $\mathsf{P}(X_{n+1} = j | X_n = i) = \mathsf{P}(X_1 = j | X_0 = i)$.

Цепь Маркова с непрерывным временем — случайный процесс $\{X_t\}_{t\geq 0}$:

$$\mathsf{P}(X_{t+h} = x_{t+h} | X_s = x_s, \quad 0 < s \leq t) =$$

$$\mathsf{P}(X_{t+h} = x_{t+h} | X_t = x_t).$$

Однородная:
$\mathsf{P}(X_{t+h} = x_{t+h} | X_t = x_t) = \mathsf{P}(X_h = x_h | X_0 = x_0)$.

Q — множество запросов, D — множество документов.

Q — множество запросов, D — множество документов.

Для некоторого подмножества $X \subset Q \times D$ известно идеальное ранжирование:

$$q_1 : \quad d_1^1 > d_2^1 > \ldots > d_{i_1}^1;$$

$$q_2 : \quad d_1^2 > d_2^2 > \ldots > d_{i_2}^2;$$

$$\ldots$$

$$q_n : \quad d_1^n > d_2^n > \ldots > d_{i_n}^n.$$

Q — множество запросов, D — множество документов.

Для некоторого подмножества $X \subset Q \times D$ известно идеальное ранжирование:

$$q_1 : \quad d_1^1 > d_2^1 > \ldots > d_{i_1}^1;$$

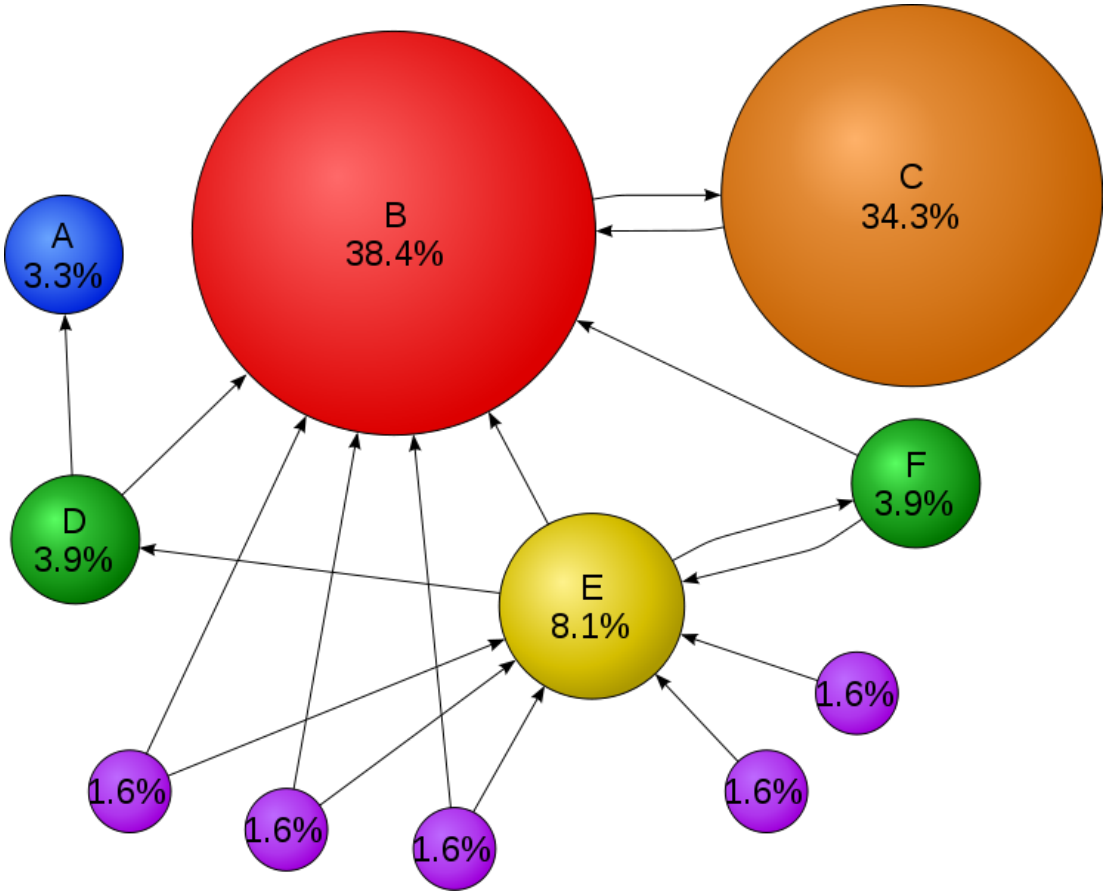
$$q_2 : \quad d_1^2 > d_2^2 > \ldots > d_{i_2}^2;$$

$$\ldots$$

$$q_n : \quad d_1^n > d_2^n > \ldots > d_{i_n}^n.$$

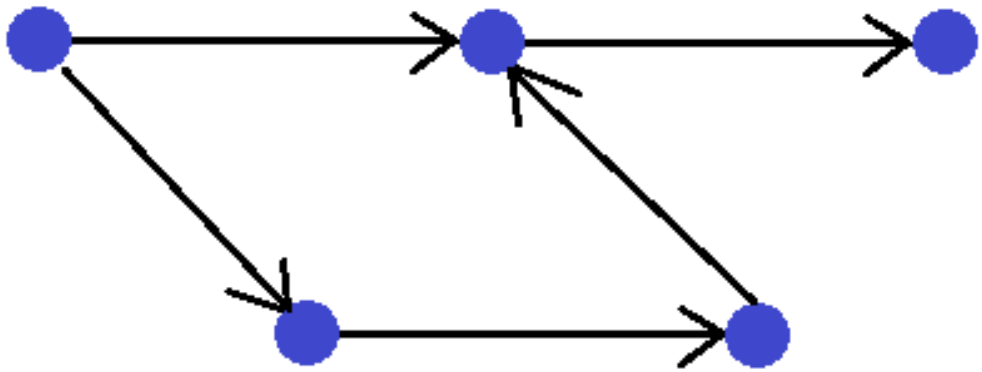
Задача: отранжировать все остальные пары из $Q \times D$.

Идея PageRank



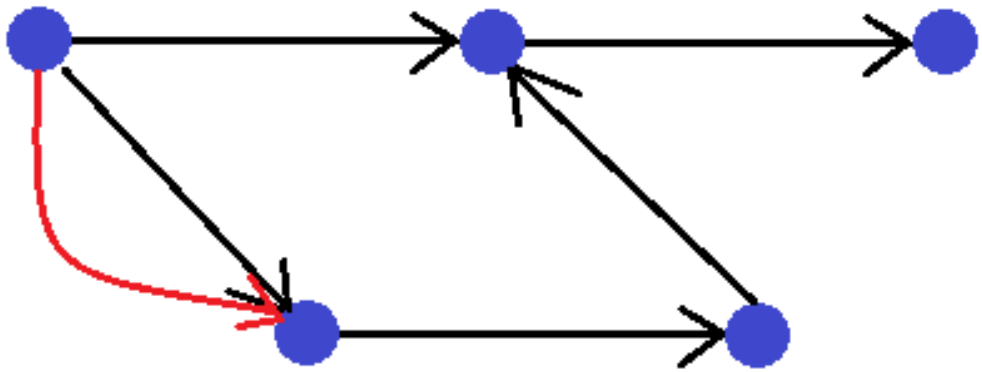
Пусть N — количество страниц в интернете.

$$\mathrm{PR}(p) = \frac{0.15}{N} + 0.85 \sum_{\tilde{p} \to p} \frac{1}{\#\{p' : \tilde{p} \to p'\}} \mathrm{PR}(\tilde{p}).$$



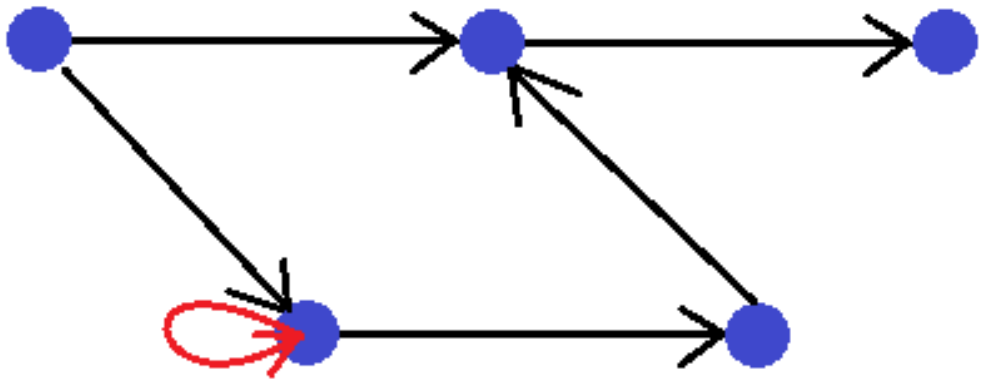
Пусть N — количество страниц в интернете.

$$\mathrm{PR}(p) = \frac{0.15}{N} + 0.85 \sum_{\tilde{p} \to p} \frac{1}{\#\{p' : \tilde{p} \to p'\}} \mathrm{PR}(\tilde{p}).$$



Пусть N — количество страниц в интернете.

$$\mathrm{PR}(p) = \frac{0.15}{N} + 0.85 \sum_{\tilde{p} \to p} \frac{1}{\#\{p' : \tilde{p} \to p'\}} \mathrm{PR}(\tilde{p}).$$



Степенной метод

$$\pi_{i+1} = A\pi_i,$$

$$A = (a_{i,j}), \quad a_{i,j} = \frac{0.15}{N} + \frac{0.85}{\text{outdegree}(i)};$$

Степенной метод

$$\pi_{i+1} = A\pi_i,$$

$$A = (a_{i,j}), \quad a_{i,j} = \frac{0.15}{N} + \frac{0.85}{\text{outdegree}(i)};$$

Метод Немировского–Нестерова

$$\pi_N = \frac{0.15}{1 - 0.85^{N+1}} \sum_{i=0}^{N} 0.85^i \tilde{A}^i \pi_0,$$

$$\tilde{A} = (\tilde{a}_{i,j}), \quad \tilde{a}_{i,j} = \frac{0.85}{\text{outdegree}(i)},$$

$$\pi_0(i) = \frac{0.15}{N}.$$

Страницы ранжируются (от самой авторитетной) по убыванию PageRank

Страницы ранжируются (от самой авторитетной) по убыванию PageRank

Плюсы

- Зависит от линковой структуры Интернета
- Согласуется с интуицией о блуждании пользователя на вебграфе
- Коррелирует с популярностью сайтов

Страницы ранжируются (от самой авторитетной) по убыванию PageRank

Плюсы

- Зависит от линковой структуры Интернета
- Согласуется с интуицией о блуждании пользователя на вебграфе
- Коррелирует с популярностью сайтов

Минусы

- Не зависит от пользовательских запросов
- Дискретное время
- Не учитывает никаких свойств Интернета кроме линковой структуры
- Не зависит от пользовательского поведения

Другие способы ранжирования с помощью марковских цепей

- Personalised PageRank
  - Зависит от пользовательских запросов
  - Дискретное время
  - Не учитывает никаких свойств Интернета кроме линковой структуры
  - Не зависит от пользовательского поведения

Другие способы ранжирования с помощью марковских цепей

- Personalised PageRank
  - Зависит от пользовательских запросов
  - Дискретное время
  - Не учитывает никаких свойств Интернета кроме линковой структуры
  - Не зависит от пользовательского поведения
- BrowseRank
  - Зависит от пользовательского поведения
  - Неперывное время
  - Не зависит от пользовательских запросов
  - Не зависит от истории пользовательского поведения
  - Не учитывает никаких свойств Интернета кроме линковой структуры и пользовательского поведения

Другие способы ранжирования с помощью марковских цепей

- Personalised PageRank
  - Зависит от пользовательских запросов
  - Дискретное время
  - Не учитывает никаких свойств Интернета кроме линковой структуры
  - Не зависит от пользовательского поведения
- BrowseRank
  - Зависит от пользовательского поведения
  - Неперывное время
  - Не зависит от пользовательских запросов
  - Не зависит от истории пользовательского поведения
  - Не учитывает никаких свойств Интернета кроме линковой структуры и пользовательского поведения
- High-Order PageRank
  - Зависит от истории пользовательского поведения
  - Не зависит от пользовательских запросов
  - Дискретное время
  - Не учитывает никаких свойств Интернета кроме линковой структуры и пользовательского поведения

Другие способы ранжирования с помощью марковских цепей

- Personalised PageRank
  - Зависит от пользовательских запросов
  - Дискретное время
  - Не учитывает никаких свойств Интернета кроме линковой структуры
  - Не зависит от пользовательского поведения
- BrowseRank
  - Зависит от пользовательского поведения
  - Неперывное время
  - Не зависит от пользовательских запросов
  - Не зависит от истории пользовательского поведения
  - Не учитывает никаких свойств Интернета кроме линковой структуры и пользовательского поведения
- High-Order PageRank
  - Зависит от истории пользовательского поведения
  - Не зависит от пользовательских запросов
  - Дискретное время
  - Не учитывает никаких свойств Интернета кроме линковой структуры и пользовательского поведения
- Supervised PageRank
  - Учитывает любые свойства страниц и ссылок в Интернете
  - Дискретное время
  - Не зависит от пользовательских запросов
  - Не зависит от истории пользовательского поведения