Введение в Data Science
Занятие 7. Ноунейм

Николай Анохин Михаил Фирулик

18 апреля 2014 г.

Работа в группе

Задача. Оценить, какой вклад внес в общий результат каждый участник группы

Шаг 1. Каждый студент **анонимно и независимо** распределяет 100 очков между всеми участниками своей группы в зависимости того, какую пользу (по его/её мнению) каждый из участников принес

Пример.

Студент	Вклад
Геральт	50
Лютик	10
Мильва	20
Регис	20

Шаг 2. Из всех оценок вычисляется общая аггрегированная оценка на основе алгоритма PageRank

План занятия

PageRank

Задача модуля

Жизнь до Google

1. Поисковые роботы используются для парсинга интернет-страниц
2. Составляется обратный индекс, в котором каждому слову соответствовал набор страниц
3. Слова из поискового запроса пользователя используются для поиска страниц в индексе
4. Из **близких** к запросу страниц формируется выдача

Проблема: Term Spam

Что придумали парни из Google

Дополнительно

1. Страницы ранжируются в соответствии с их “важностью” с помощью алгоритма PageRank
2. О релевантности страниц судят не только по словам, находящимся на текущей странице, но и по словам “соседних” страниц

Random Surfer

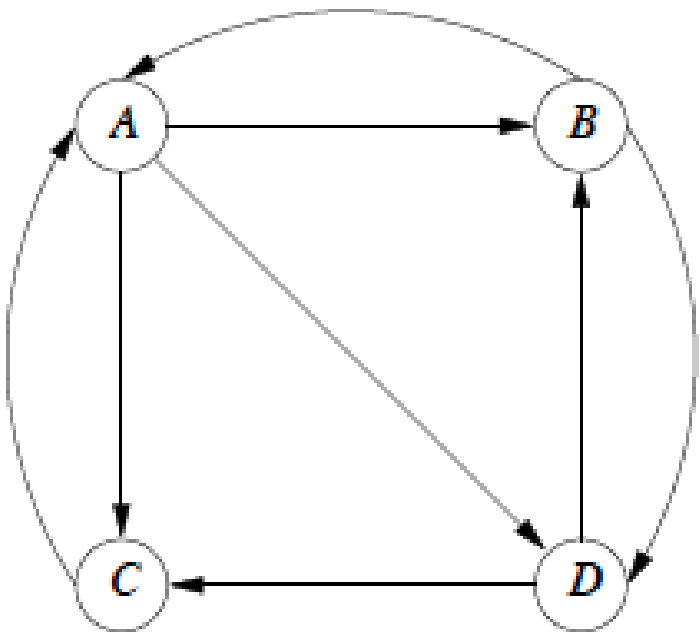
Интуиция

Пользователь начинает с просмотра случайной страницы, после чего с равной вероятностью переходит по одной из ссылок на этой странице. Процесс продолжается до бесконечности. PageRank страницы – вероятность обнаружить пользователя на этой странице.

- Пользователь с большей вероятностью посещает “полезные” страницы, чем “бесполезные”
- Создатели страниц размещают ссылки на “полезные” страницы

PageRank

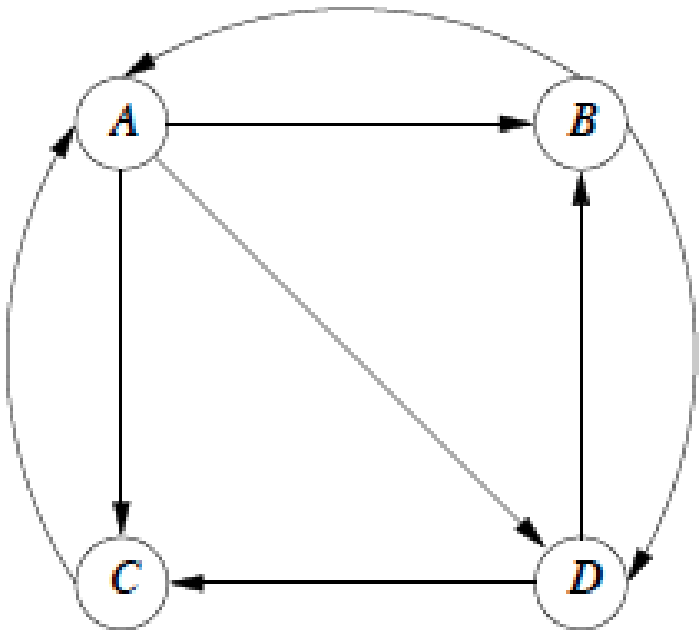
Представим интернет, как направленный граф со страницами в качестве вершин и ссылками между страницами в качестве ребер



Матрица вероятностей перехода

$$M = \begin{pmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{pmatrix}$$

PageRank



Элементы матрицы перехода

$$m_{ij} = P(\mathbf{v}_i^{(k)} | \mathbf{v}_j^{(k-1)})$$

Изначально все страницы равновероятны

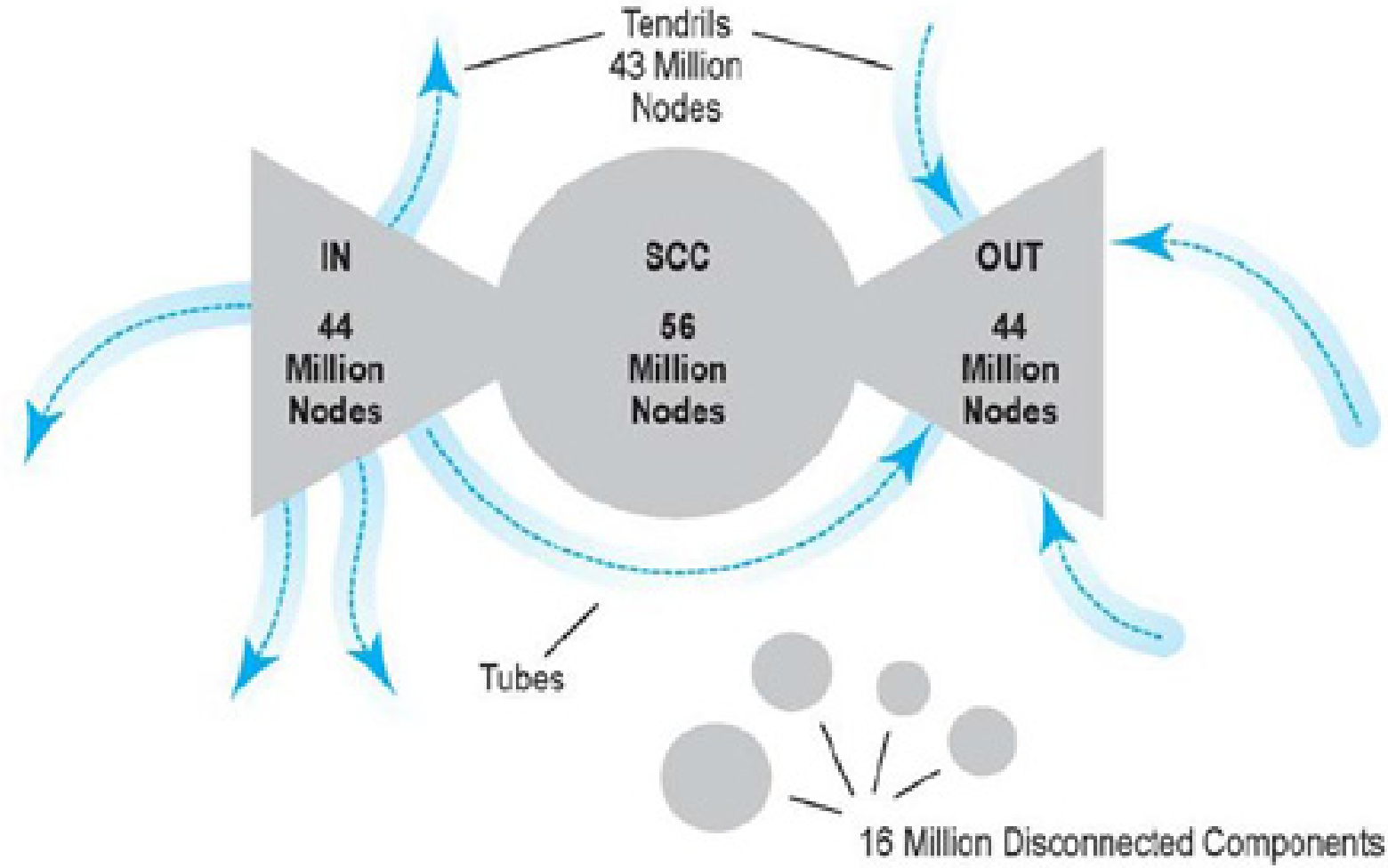
$$\mathbf{v}^{(0)} = \begin{pmatrix} 1/n & \dots & 1/n \end{pmatrix}^\top$$

Вектор вероятностей на k шаге

$$\mathbf{v}^{(k)} = M\mathbf{v}^{(k-1)}$$

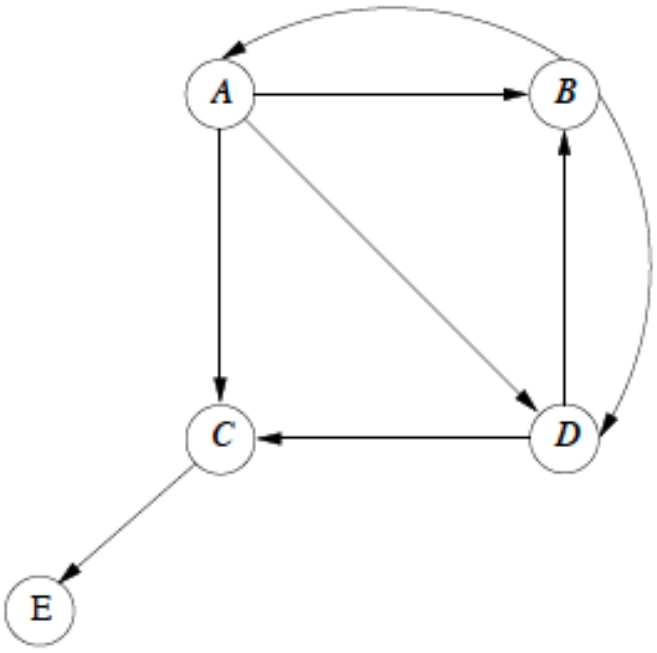
Предельное значение $\mathbf{v}$ – собственный вектор M, соответствующий собственному числу $\lambda = 1$. Процесс сходится, если из любой вершины можно попасть в любую.

Структура Интернета

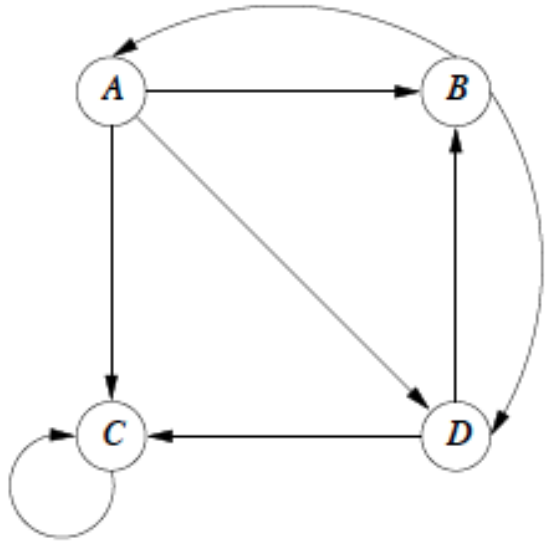


Проблемы PageRank

Dead End



Spider Trap



Решение. разрешим пользовалю “телепортироваться” на случайную страницу с вероятностью $1 - \beta$

$$\mathbf{v}^{(k)} = \beta M \mathbf{v}^{(k-1)} + (1 - \beta)\frac{\mathbf{e}}{n}$$

Пример

Матрица перехода

$$M = \begin{pmatrix} 0 & 1/2 & 0 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 1 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{pmatrix}$$

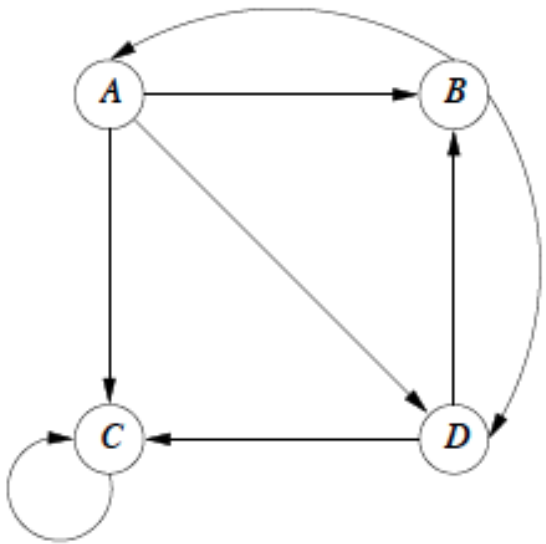
Без телепортов

$$\mathbf{v} = \begin{pmatrix} 0 & 0 & 1 & 0 \end{pmatrix}$$

С телепортами $\beta = 0.8$

$$\mathbf{v} = \begin{pmatrix} \frac{15}{148} & \frac{19}{148} & \frac{95}{148} & \frac{19}{148} \end{pmatrix}$$

Spider Trap



Методика оценки

	Геральт	Лютик	Мильва	Регис	Индивидуально
Геральт	50	10	30	30	20
Лютик	10	70	10	5	5
Мильва	20	10	30	30	15
Регис	20	10	30	35	15

Матрица перехода, $\beta = 0.9$

$$M = \begin{pmatrix} 0.5 & 0.1 & 0.3 & 0.3 \\ 0.1 & 0.7 & 0.1 & 0.05 \\ 0.2 & 0.1 & 0.3 & 0.3 \\ 0.2 & 0.1 & 0.3 & 0.35 \end{pmatrix} \quad v = \begin{pmatrix} 0.31 \\ 0.23 \\ 0.23 \\ 0.24 \end{pmatrix}$$

Групповая оценка: 30/40

Итог:

Геральт: 29, Лютик: 12, Мильва: 22, Регис: 22