ЦИФРОВАЯ КУЛЬТУРА
УНИВЕРСИТЕТ ИТМО

PageRank с использованием GraphBLAS

Рустам Азимов

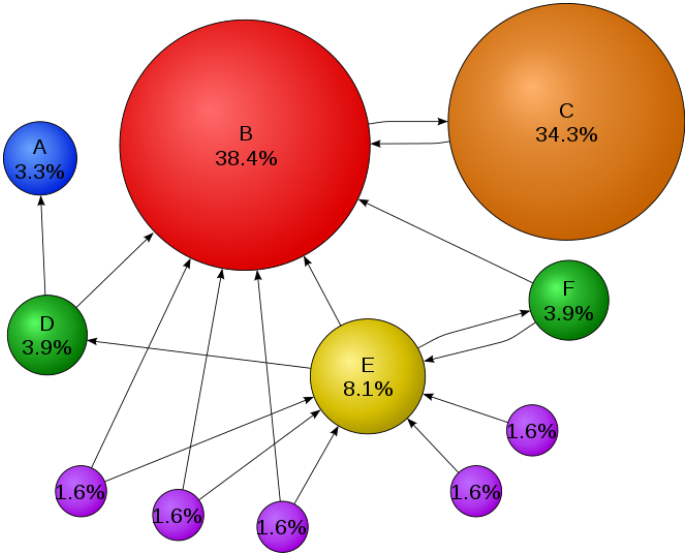
PageRank

- **PageRank** — один из алгоритмов ссылочного ранжирования, который применяется к коллекции документов, связанных гиперссылками
- В результате каждому документу назначается некоторое численное значение, измеряющее его «важность» или «авторитетность» среди остальных документов

Связь с графами

- Если представить интернет как набор веб-страниц, в каждой из которых могут быть ссылки на другие страницы, то такие связи легко описываются с помощью графа
- Вершины — веб-страницы
- Рёбра — ссылки
- Граф ориентированный?
- PageRank может применяться не только к веб-страницам, но и к любому набору объектов, связанных между собой взаимными ссылками, то есть к любому графу

Пример



Особенности графа

- Большой размер
- Разреженность
- Наличие петель
- Наличие циклов и обратных рёбер

Идея

- Чем больше ссылок у странцы, тем она важнее
- Какие ссылки считаем исходящие или входящие?
- Все ли ссылки одинаково важны?

Идея

- Чем важнее страница, тем важнее исходящая из неё ссылка
  - Вклад каждой ссылки должен быть пропорционален важности страницы, от которой она исходит
  - Если страница i с важностью r_i имеетd_i ссылок, то каждая исходящая ссылка будет давать вклад $\frac{r_i}{d_i}$
  - Страница j будет иметь важность, равную сумме входящих ссылок

Связь с линейной алгеброй

- Рассмотрим стохастическую матрицу смежности
- Если есть ссылка из i в j, то $M_{ij} = \frac{1}{d_i}$
- Тогда записать описанную идею можно в матричном виде

$$r = M \otimes r$$

Суть процесса

- Поймем суть матричной формы $r = M \otimes r$
- Представим пользователя, блуждающего по сети Интернет
- В момент времени t он попадает на страницу i
- В момент времени $t + 1$ он следует по случайной ссылке со страницы i
- Он оказывается на странице j, попав на нее из i
- Процесс продолжается бесконечно
- Пусть $p(t)$ — вектор, у которого на i-ом месте будет вероятность того, что пользователь окажется на странице i в момент времени t
- Тогда $p(t)$ — вероятностное распределение на всех страницах

Суть процесса

- Движение по ссылкам случайным равновероятностным образом можно описать как
$$p(t+1) = M \otimes p(t)$$
- Представим, что в какой-то момент мы достигнем неподвижной точки
$$p(t+1) = M \otimes p(t) = p(t)$$
- Тогда $p(t)$ — стационарное распределение случайного блуждания
- Получается из $r = M \otimes r$ следует, что r — стационарное распределение

Power Iteration

- Инициализация — $r^0 = [1/N, ..., 1/N]$
- Шаг $r^{t+1} = M \otimes r^t$
- Повторять, пока не $|r^{t+1} - r^t|_1 < \varepsilon$
- Какие возникнут проблемы?

Power Iteration

- Инициализация — $r^0 = [1/N, ..., 1/N]$
- Шаг $r^{t+1} = M \otimes r^t$
- Повторять, пока не $|r^{t+1} - r^t|_1 < \varepsilon$
- Какие возникнут проблемы?
  - тупики

Power Iteration

- Инициализация — $r^0 = [1/N, ..., 1/N]$
- Шаг $r^{t+1} = M \otimes r^t$
- Повторять, пока не $|r^{t+1} - r^t|_1 < \varepsilon$
- Какие возникнут проблемы?
  - тупики
  - циклы
- Как решить эти проблемы?

Суть процесса

- Вводим телепорт
- На каждом шаге с вероятностью β пользователь выбирает одну из d_i ссылок на странице
- А с вероятностью $1 - \beta$ телепортируется на случайную страницу
- Обычно $\beta = 0.85$
- За конечное число шагов пользователь выпрыгнет из цикла
- Из тупиков всегда телепортируемся

PageRank: итерация

$$r_j = \sum_{i \to j} \beta \frac{r_i}{d_i} + (1 - \beta) \frac{1}{N}$$

PageRank: преобразование матрицы

$$G = \beta\, M + (1 - \beta) \left[\frac{1}{N}\right]_{N \times N}$$