PageRank

Тимур Исхаков

Высшая школа экономики, Факультет компьютерных наук

16 мая 2017

Требуется упорядочить объекты.

Желательно, присвоив объектам некие коэффициенты.

Приложения

- Ранжирование документов в поиске (search engine results page)
- Предсказание количества ссылок на документ
- Ранжирование товаров, пользователей (например, в Twitter)
- Лексическая семантика
- Подсчет влияния (impact)
- ...

Рассмотрим интернет как ориентированный граф, где вершина соответствует странице, а ребро — гиперссылке.

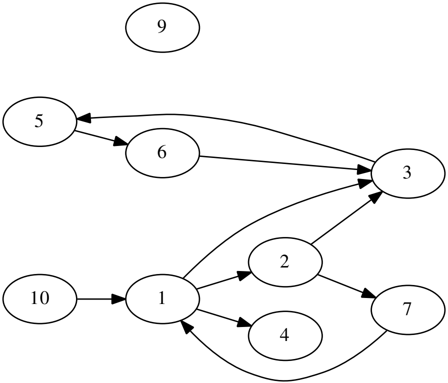


Рис.: Пример ориентированного графа

Представим себе случайное блуждание пользователя по страницам в интернете.

Зададим H — матрицу переходов: пусть из страницы i выходит l_i ссылок, тогда

$$H_{i,j} := \begin{cases} \frac{1}{l_i}, & \text{если есть ориентированное ребро } (i,j); \\ 0, & \text{иначе.} \end{cases}$$

n — количетсво страниц.

Посмотрим на «наивный» PageRank:

$$R : \text{страница} \to \mathbb{R},$$

такая что верно равенство:

$$R(u) = \sum_{v \in \delta_{in}(u)} \frac{R(v)}{l_v}$$

$$\pi : (R(1), \ldots, R(n))$$

$$\pi^T = \pi^T \cdot H$$

Степенной алгоритм:

$$\pi^{(k)^T} = \pi^{(k-1)^T} H$$

- Сходится ли он?
- Насколько долго сходится?
- Зависит ли сходимость от H, $\pi^{(0)}$, каких-то параметров?
- Можно ли как-то интерпретировать полученное π?

Описанный нами процесс блуждания очень напоминает движение по Марковской цепи.

Однако, заданная матрица не является *стохастической* из-за «тупиковых» вершин (*dangling nodes*).

Исправим проблему, считая, что из «тупиковой» вершины пользователь равновероятно идет в любую страницу:

$$S := H + \left(1/n\, a\, \mathbb{1}^T\right), \text{ где } a_i := [l_i = 0]$$

Предложение

Стохастическая квадратная матрица A имеет собственное значение 1.

Рассмотрим неравенство $(A - I)x \geqslant b, b > 0$. Пусть оно имеет решение x^*, тогда $Ax^* \geqslant x^* + b$. Но каждый элемент Ax — выпуклая комбинация элементов x, следовательно не превосходит x^*_{max}. Но хотя бы один элемент $x^* + b$ больше, чем x^*_{max}. Следовательно, неравенство не имеет решений.

Предложение

Стохастическая квадратная матрица A имеет собственное значение 1.

Неравенство $(A - I)x \geqslant b, b > 0$ не имеет решений.
Тогда рассмотрим задачу LP:

$$\begin{cases} (A - I)y \geqslant 1 \\ 0^T y \to \min \end{cases}$$

Она несовместна, следовательно двойственная несовместна или неограничена.

$$\begin{cases} x^T(A - I) = 0^T \\ x \geqslant 0 \\ \mathbb{1}^T x \to \max \end{cases}$$

$x = 0$ является решением, следовательно существует невырожденное решение.

Циклы, в которые входят ребра, но из которых не выходит ребер, бесконечно ракапливают $R(u)$ (в статье ситуацию называют rank sink).

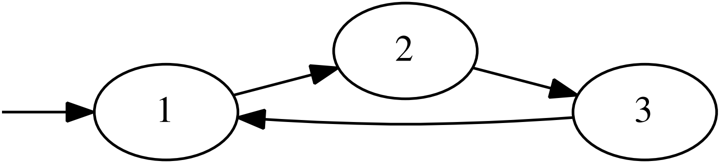


Рис.: Rank sink

Введем дополнительный вариант действия пользователя: потерять интерес к текущей странице и перейти на совсем другую. Пусть он переходит на другие страницы равновероятно.

$$G := \alpha S + (1-\alpha) 1/n \, \mathbb{1}\mathbb{1}^T, \; \alpha \in (0,1)$$

α называют *damping factor*, обычно берут равным 0.85.

Google matrix

$$G := \alpha S + (1-\alpha) 1/n \, \mathbb{1}\mathbb{1}^T, \ \alpha \in (0,1)$$

Заметим, что:

- G — стохастическая матрица (stochastic): $\sum_i G_i = 1$;
- G — неприводимая матрица (irreducible): $G_{i,j} > 0$
- G — апериодическая матрица (aperiodic): неприводимость является достаточным условием
- G — примитивная матрица (primitive) $\exists\, k : \ G^k > 0$

$$G_{i,j} > 0$$

Теорема (Perron–Frobenius)

Если квадратная матрица A с вещественными элементами положительна, то

- $$\min_i \sum_j A_{i,j} \leqslant \lambda \leqslant \max_i \sum_j A_{i,j}$$
- $$\max \lambda \text{ входит с кратностью } 1$$

Таким образом, $\max \lambda = 1$ и входит с кратностью 1.

Добавив ограничения $\|\pi\|_1 = 1$, $\pi \geqslant 0$, получим единственность.

При этом π также будет неким распределением.

Способы вычисления π:

Степенной метод:

$$\begin{cases} \pi^T - \pi^T G \\ \|\pi\|_1 = 1 \end{cases}$$

Линейная система:

$$\begin{cases} \pi^T (I - G) = 0^T \\ \|\pi\|_1 = 1 \end{cases}$$

$$\begin{aligned} {\pi^{(k)}}^T &= {\pi^{(k-1)}}^T G \\ &= \alpha {\pi^{(k-1)}}^T S + \frac{1-\alpha}{n} {\pi^{(k-1)}}^T \mathbb{1}\mathbb{1}^T \\ &= \alpha {\pi^{(k-1)}}^T H + \left(\alpha {\pi^{(k-1)}}^T a + 1 - \alpha\right) \mathbb{1}^T / n \end{aligned}$$

Скорость сходимости степенного метода равна скорости сходимости $\left|\frac{\lambda_2}{\lambda_1}\right|$ к 0.
$\lambda_1 = 1$, следовательно сходимость зависит от λ_2.

Предложение

Пусть S — стохастическая матрица со спектром $\{1, \lambda_2, \ldots\}$. Тогда $G = \alpha S + (1 - \alpha)\mathbb{1}v^T$ имеет спектр $\{1, \alpha\lambda 2, \ldots\}$.

На практике λ_2 близко к 1, поэтому скорость сходимости напрямую зависит от значения α.

Предложение

Пусть S — стохастическая матрица со спектром $\{1, \lambda_2, \ldots\}$. Тогда $G = \alpha S + (1-\alpha)\mathbb{1}v^T$ имеет спектр $\{1, \alpha\lambda 2, \ldots\}$.

S — стохастическая, $(\lambda = 1, e)$ — собственная пара. Пусть $Q = \begin{pmatrix} e & X \end{pmatrix}$ — невырожденная матрица. $Q^{-1} = \begin{pmatrix} y^T \\ Y^T \end{pmatrix}$.

Тогда $Q^{-1}Q = \begin{pmatrix} y^Te & y^TX \\ Y^Te & Y^TX \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & I \end{pmatrix}$. $y^Te = 1$, $Y^Te = 0$.

$$Q^{-1}SQ = \begin{pmatrix} y^Te & y^TSX \\ Y^Te & Y^TSX \end{pmatrix} = \begin{pmatrix} 1 & y^TSX \\ 0 & Y^TSX \end{pmatrix}$$

Значит, спектр Y^TSX — $\{\lambda_2, \ldots\}$.

Предложение

Пусть S — стохастическая матрица со спектром $\{1, \lambda_2, \dots\}$. Тогда $G = \alpha S + (1-\alpha)1v^T$ имеет спектр $\{1, \alpha\lambda 2, \dots\}$.

Спектр $Y^T SX$ — $\{\lambda_2, \dots\}$.

$$\begin{aligned}
Q^{-1}GQ &= \alpha Q^{-1}SQ + (1-\alpha)Q^{-1}\mathbb{1}v^T Q \\
&= \begin{pmatrix} \alpha & \alpha y^T SX \\ 0 & \alpha Y^T SX \end{pmatrix} + (1-\alpha)\begin{pmatrix} y^T\mathbb{1} \\ Y^T\mathbb{1} \end{pmatrix}\begin{pmatrix} v^T e & v^T X \end{pmatrix} \\
&= \begin{pmatrix} \alpha & \alpha y^T SX \\ 0 & \alpha Y^T SX \end{pmatrix} + \begin{pmatrix} 1-\alpha & (1-\alpha)v^T X \\ 0 & 0 \end{pmatrix} \\
&= \begin{pmatrix} 1 & \alpha y^T SX + (1-\alpha)v^T X \\ 0 & \alpha Y^T SX \end{pmatrix}
\end{aligned}$$

Таким образом, спектр G — $\{1, \alpha\lambda_2, \dots\}$.

Предложение

Пусть x — решение системы $x^T(I - \alpha H) = v^T$. Тогда $\pi = x/\|x\|_1$.

$$
\begin{aligned}
1 &= v^T \mathbb{1} \\
&= x^T(I - \alpha H)\mathbb{1} \\
&= x^T \mathbb{1} - \alpha x^T H \mathbb{1} \\
&= x^T \mathbb{1} - \alpha x^T (\mathbb{1} - a) \\
&= (1 - \alpha) x^T \mathbb{1} + \alpha x^T a
\end{aligned}
$$

Предложение

Пусть x — решение системы $x^T(I - \alpha H) = v^T$. Тогда $\pi = x/\|x\|_1$.

$$(1-\alpha)x^T\mathbb{1} + \alpha x^T a = 1$$

$$\begin{aligned} x^T(I - G) &= x^T\left(I - \alpha H - \alpha a v^T - (1-\alpha)\mathbb{1}v^T\right) \\ &= x^T(I - \alpha H) - x^T\left(\alpha a + (1-\alpha)\mathbb{1}\right)v^T \\ &= v^T - v^T = 0^T \end{aligned}$$

И также $\|x/\|x\|_1\|_1 = 1$.

α

α	Количество операций
.5	34
.75	81
.8	104
.85	142
.9	219
.95	449
.99	2 292
.999	23 015

При α близком к 1, π чувствительно к небольшим изменениям α.

Матрица переходов H

- Можно заполнять не равновероятно, а собирать информацию с пользователей.
- Кнопка «назад»

Для марковской цепи с матрицей переходов P π чувствительно к перестановкам в P т. и т. т., когда $|\lambda_2(P)| \approx 1$. $|\lambda_2(P)| \leqslant \alpha$.

v

Немного изменим нашу матрицу:

$$+(1-\alpha)1/n\mathbb{1}\mathbb{1}^T \to +(1-\alpha)\mathbb{1}v^T$$

$$\pi(v)^{(k)}{}^T = \ldots = \alpha\pi(v)^{(k-1)}{}^T H + \left(\alpha\pi(v)^{(k-1)}{}^T a + 1 - \alpha\right) v^T$$

v

Немного изменим нашу матрицу:

$$+(1-\alpha)1/n\mathbb{1}\mathbb{1}^T \to +(1-\alpha)\mathbb{1}v^T$$

$$\pi(v)^{(k)^T} = \ldots = \alpha\pi(v)^{(k-1)^T}H + \left(\alpha\pi(v)^{(k-1)^T}a + 1 - \alpha\right)v^T$$

Персонализированное ранжирование:

$$\pi(user) = \sum_i \beta(user)_i \pi(v_i)$$

- В случае, когда H задает случайное блуждание, можно заменить H на $D^{-1}L$, где $D^{-1}_{i,i} = \frac{1}{l_i}$, а L — матрица смежности.
  Таким образом, вычисление $\pi^T H$ заменится на $\pi^T D^{-1} L = \pi^T \operatorname{Diag}\{D^{-1}\} \times L$, что уменьшает количество умножений с $\operatorname{nnz}(H)$ до n.
- Можно «сжать» тупиковые вершины в одну, а затем аккуратно восстановить их ранг.
- Нам не обязательно сходиться в π, так как важен в основном порядок страниц.
- ...

Обновление

Все плохо.

Связи можно обновить. Но за $\mathcal{O}(n^3)$.

Добавить страницы дельзя.

Можно начать, используя предыдущий π. Но сильно быстрее не будет.

Обновление

Аппроксимация. Не в этот раз.

Exact aggregation, approximate aggregation.

Можно делить на блоки, считать приближенно.

Christian Borgs, Michael Brautbar, Jennifer Chayes, Shang-Hua Teng, 2012

Given an arbitrary approximation factor $c > 1$, to output a set S of nodes that with high probability contains all nodes of PageRank at least Δ, and no node of PageRank smaller than $\frac{\Delta}{c}$. We call this problem SIGNIFICANTPAGERANKS. We develop a nearly optimal, local algorithm for the problem with runtime complexity $\Omega\left(\frac{n}{\Delta}\right)$ on networks with n nodes. We show that any algorithm for solving this problem must have runtime of $\Omega\left(\frac{n}{\Delta}\right)$, rendering our algorithm optimal up to logarithmic factors.

TrustRank

У PageRank есть довольно важная проблема: восприимчивость к спаму.

Пусть у нас есть оракул $O(p) = \begin{cases} 1, & \text{если p — «хорошая»}, \\ 0, & \text{если p — «плохая»}. \end{cases}$

В идеале, мы хотели бы знать «надежность» страницы $T(p) = \Pr[O(p) = 1]$. Однако, такое T непонятно, как получить. Уменьшим требования к T. Варианты для подсчета качества:

1. $$T(p) < T(q) \Leftrightarrow \Pr[O(p) = 1] < \Pr[O(q) = 1]$$
   $$T(p) = T(q) \Leftrightarrow \Pr[O(p) = 1] = \Pr[O(q) = 1]$$
2. $$T(p) > \delta \Leftrightarrow O(p) = 1$$

Возьмем подвыборку страниц $\mathcal{S}$, вычислим для них значения O.

Если $O(p) = 1$, разумно предполагать, что она ссылается на «хорошие» страницы. Однако, по мере удаления от p уверенность в этом падает.

Можно использовать два подхода: угасание доверия $(1 \to \beta \to \beta^2 \to \dots)$ и равномерно распределение доверия (есть связь из p в v, тогда $O(v) = \frac{O(p)}{l_p}$).

Когда в вершину идет несколько связей, для первого случая логично использовать максимум, для втого сумму (и добавить нормировку).

Осталось выбрать подвыборку страниц $\mathcal{S}$.

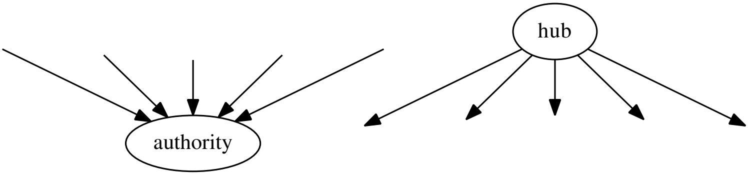
Нам хотелось бы взять $\mathcal{S}$ размера L так, чтобы $|d(\mathcal{S})|$ (количество прямых соседей) было как можно больше. Однако, это NP-трудная задача.
Эвристики:

1. Первые L страниц для PageRank.
2. Первые L страниц для «инвертированного» PageRank.

Сам TrustRank можно использовать для отсеивания страниц при вычисления PageRank.

HITS (Hyperlink Induced Topic Search)

PageRank	HITS
Одно значение на страницу	Два значения: authorities x_i и hubs y_i
Не зависит от запроса	Зависит от запроса



Authority — страница, у которой много входящих ребер,
Hub — страница, у которой много выходящих ребер.

Good authorities are pointed to by good hubs and good hubs point to good authorities.

x_i — authority score, y_i — hub score для страницы i.
E — множество направленных ребер веб-графа.

$$x_i^{(k)} = \sum_{j \in \delta_{in}(i)} y_j^{(k-1)}$$

$$y_i^{(k)} = \sum_{j \in \delta_{out}(i)} x_j^{(k)}$$

L – матрица смежности. Тогда:

$$x^{(k)} = L^T y^{(k)}$$

$$y^{(k)} = Lx^{(k-1)}$$

Или:

$$x^{(k)} = L^T L x^{(k-1)}$$
$$y^{(k)} = L L^T y^{(k-1)}$$

HITS (Hyperlink Induced Topic Search)

1. Запрос → **neighborhood graph N**
   - Страницы, содержащие слова запроса (**inverted file index**)
   - N расширяется добавлением вершин, которые указывают на N и на которые указывает N (+ синонимы)
2. Вычислиить x и y для каждой страницы в N
   - $N \to L$
   - Собственный вектор $L^T L \to x, \; y = Lx$

SALSA (Stochastic Approach for Link-Structure Analysis)

Вместо того, чтобы строить в HITS граф L, построим неориентированный двудольный граф G с долями V_h (для всех страниц с исходящими ссылками) и V_a (для всех страниц со входящими ссылками). Множество ребер строится по ребрам в графе N.

По графу G строятся матрицы Марковских цепей для A и H, соответственно π_a и π_h задают ранги для authorities и hubs (отдельно в каждой компоненте сильной связности).