Ссылочный граф Интернета

Лекция 6

БГУ ФПМИ, 2018

План

Структура и размер веб-графа

PageRank

Ссылочный спам

План

Структура и размер веб-графа

PageRank

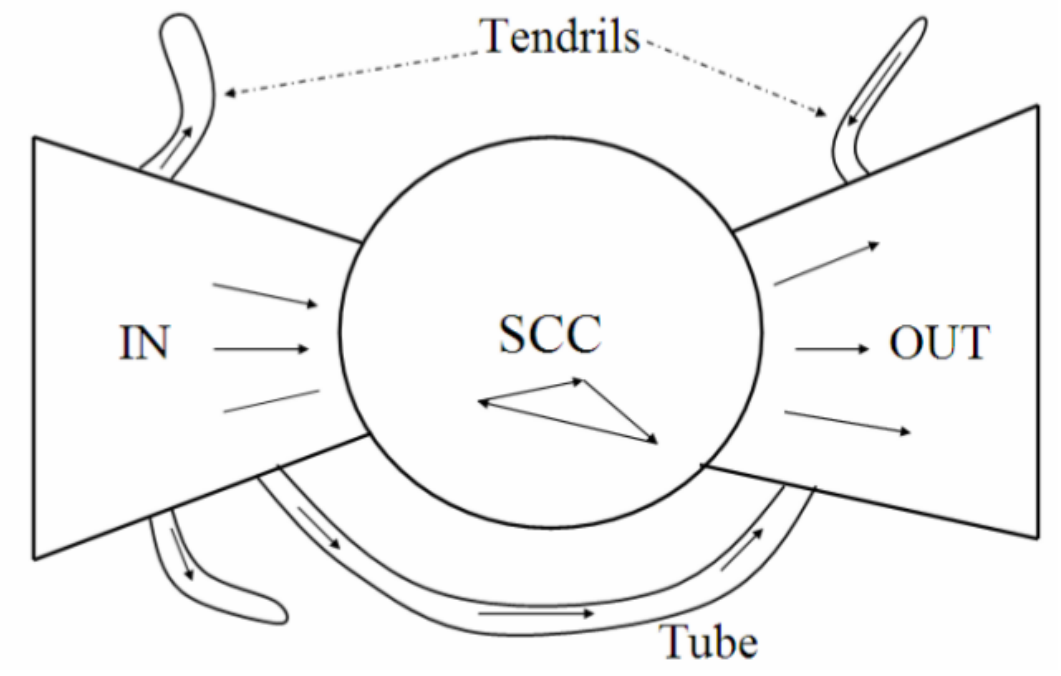
Ссылочный спам

Веб как граф

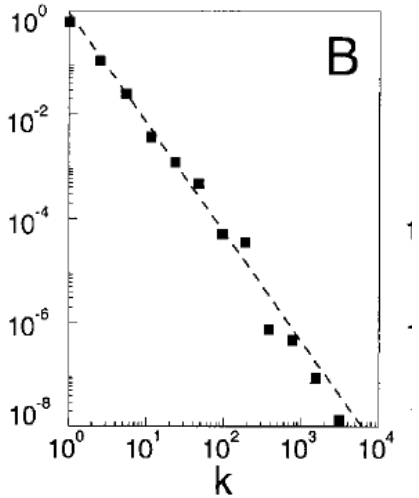
Статическую часть Интернета, состоящую из HTML-страниц и ссылок между ними, можно представить в виде ориентированного графа, в котором каждой странице соответствует вершина, а гиперссылке — ребро.

Веб как граф

Статическую часть Интернета, состоящую из HTML-страниц и ссылок между ними, можно представить в виде ориентированного графа, в котором каждой странице соответствует вершина, а гиперссылке — ребро.



Структура графа



$$p(d_v = k) = Ck^{-\gamma}$$

Для d_v^{in} — $\gamma \approx 2.1$

Такой закон распределения характерен для модели предпочтительного присоединения.

Другие свойства

- Размер наибольшей связной компоненты $\Theta(N)$
- Диаметр порядка $\log N$.
- Отрицательная корреляция степеней (disassortativity).

Размер веба

- Порядка 1 млрд. сайтов.
- Более-менее точно можно установить только количество проиндексированных страниц (порядка 47 млрд в англоязычном интернете).
- Не учтен deep web.

План

Структура и размер веб-графа

PageRank

Ссылочный спам

PageRank algorithm

Algorithm 1. (Poor man's PageRank algorithm.)

Input: Given a transition matrix P, a teleportation vector v, and a coefficient c

Output: Compute PageRank p

begin

Initialize $p^{(0)} = v$, $k = 0$

repeat

$p^{(k+1)} = cP^T p^{(k)}$

$\gamma = \|p^{(k)}\| - \|p^{(k+1)}\|$

$p^{(k+1)} = p^{(k+1)} + \gamma v$

$\delta = \|p^{(k+1)} - p^{(k)}\|$

$k = k + 1$

until $\delta < \epsilon$

end

return $p^{(k)}$

PageRank algorithm

```
1   for i ← 1 to N do
2       R[i] ← 1
3   loop
4       for i ← 1 to N do
5           R'[i] ← 1 − δ
6       for k ← 1 to E do
7           i ← link[k].from
8           j ← link[k].to
9           R'[j] ← R'[j] + δ·R[i] / out(i)
10      s ← N
11      for i ← 1 to N do
12          s ← s − R'[i]
13      for i ← 1 to N do
14          R[i] ← R'[i] + s/N
```

Extended PageRank algorithm

```
1   for i ← 1 to N do
2       R[i] ← J[i] · N
3   loop
4       for i ← 1 to N do
5           R'[i] ← (1 − δ) · J[i] · N
6       for k ← 1 to E do
7           i ← link[k].from
8           j ← link[k].to
9           R'[j] ← R'[j] + δ · R[i] · F[i, j]
10      s ← N
11      for i ← 1 to N do
12          s ← s − R'[i]
13      for i ← 1 to N do
14          R[i] ← R'[i] + s · J[i]
```

HITS

$$a(i) = w_a \sum_{j \to i} h(j)$$

$$h(i) = w_h \sum_{i \to j} a(j)$$

HITS

W – бинарная матрица смежности графа, тогда

$$a^{(k+1)} = w_a W^T h^{(k)} = w_a w_h W^T W a^{(k)}$$

$$h^{(k+1)} = w_a w_h W W^T h^{(k)}$$

$$a^{(0)} = h^{(0)} = \{1/\sqrt{N}, 1/\sqrt{N}, \dots 1/\sqrt{N}\}$$

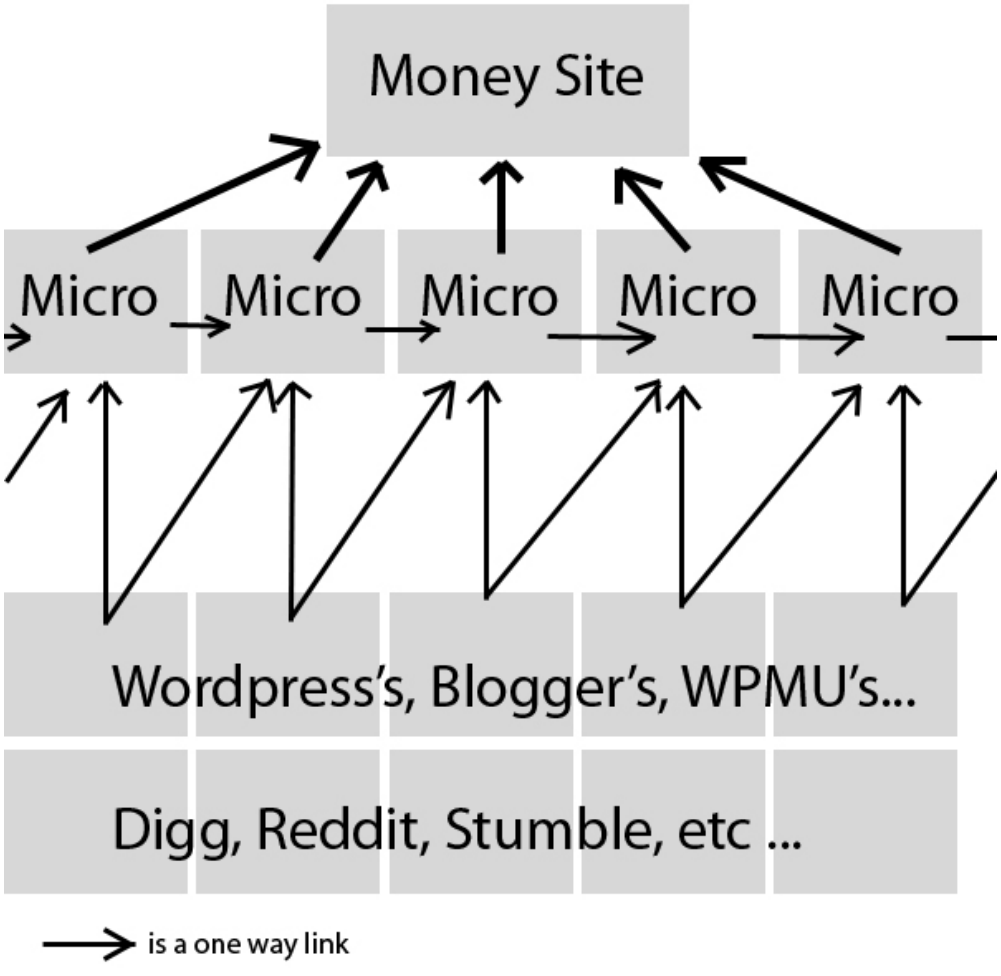
План

Структура и размер веб-графа

PageRank

Ссылочный спам

Популярность ссылочных алгоритмов ранжирования привела к появлению линковых ферм — кликообразных структур в веб-графе, каждая страница в которых нужна только лишь для того, чтобы ссылаться на другие страницы. Такие структуры могут завышать PageRank страниц, на которые из них ведут ссылки.


Money Site
Micro
Micro
Micro
Micro
Micro
Wordpress's, Blogger's, WPMU's...
Digg, Reddit, Stumble, etc ...
is a one way link

Следующая лекция

Индексация документов