

Лекция №5

Ссылочное ранжирование

Евгений Чернов

Ссылочное ранжирование



1998 год



[New](#)



[Cool](#)

YAHOO!



[Today's News](#)



[More Yahoos](#)

[1998 Winter Games](#)

results, schedules, news



**MegaMarketing
BENEFITS EXPOSED!**

[Academy Award](#)

[Nominations](#)

Search

[options](#)

[Yahoo! Chat](#) with Wall Street guru **Jim Cramer**, supermodel **Frederique**

[Yellow Pages](#) - [People Search](#) - [Maps](#) - [Classifieds](#) - [Personals](#) - [Chat](#) - [Free Email](#)

[Shopping](#) - [My Yahoo!](#) - [News](#) - [Sports](#) - [Weather](#) - [Stock Quotes](#) - [more...](#)

- [Arts and Humanities](#)

Architecture, Photography, Literature...

- [News and Media \[Xtra!\]](#)

Current Events, Magazines, TV, Newspapers...

1998 год



Google!
B E T A

Search the web using Google!

Google Search

I'm feeling lucky

PageRank!



Larry Page



Sergey Brin



Проблема: как показать vk.com по запросам:

вконтакте

вконтакте вход главная

вконтаткте

еонтаки

однаклассники вконтакте

вконтакте войти на сайт

juunfrn

как зарегистрироваться в контакте

dronтакте

контакт вход

вк добро пожаловать

www vk com

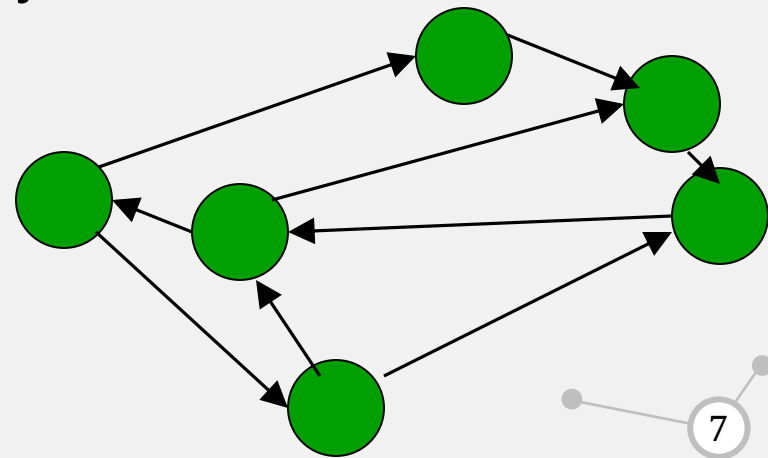
млющьюкг

вкакнтak

+ 100500 запросов



- Что есть кроме тела документа?
 - Рассмотрим ссылки между страницами
- Возникают вопросы
 - Достаточно ли ссылки авторитетны?
 - Они полезны для ранжирования?
 - Если на странице есть ссылка на CERN, то страница про ядерную физику?
- Веб – граф ссылок
- У ссылок есть [текст](#)





Единая Россия

Материал из Википедии — свободной энциклопедии

*У этого термина существуют и другие значения, см. Единая Россия (значения).
Запрос «ЕР» перенаправляется сюда; см. также другие значения.*


Всероссийская политическая партия «Еди́ная Росси́я» — официально зарегистрированная российская политическая партия, крупнейшая партия России^[4]. По итогам выборов 2003 года «Единая Россия» сформировала в Государственной думе парламентское большинство, в 2007 году — конституционное большинство, но в 2011 году утратила конституционное большинство, но сохранила абсолютное большинство. Лидер партии, возглавлявший избирательный список партии на думских выборах 2007 года — Президент Российской Федерации Владимир Путин.

Индексация анкорного текста



✕ 🔍

[Интернет](#) [Картинки](#) [Видео](#) [Приложения](#) [Новости](#) [Обсуждения](#) [Ответы](#)

 **Партия Жуликов и Воров**
[партия-жуликов-и-воров.рф](#)
Александр Хинштейн: «Лучше быть в «партии жуликов и воров», чем в «партии убийц, насильников и грабителей.»

[Лучшие 80 плакатов конкурса "Единая Россия" - партия жуликов и воров"](#)
[kprf.ru/crisis/agitator/88691.html](#)
Публикуем лучшие 80 плакатов конкурса "Единая Россия" - партия жуликов и воров", который организовал блогер Навальный.

[Единая Россия — Википедия](#)
[ru.wikipedia.org/wiki/Единая_Россия](#)
Всероссийская политическая **пáртия** «Еди́ная Росси́я» — официально зарегистрированная российская правящая политическая **партия**, крупнейшая в стране. Образована 1 декабря 2001 года в форме объединения политических движений «Единство» и «Отечество — Вся Россия». По итогам выборов 2003 года «Единая...



Индексация анкорного текста



texts [98]	text	партия жуликов и воров
	type	ExtLink
	rank	53.0
	attribute [0]	1
	attribute [1]	4487
	attribute [2]	52
	attribute [3]	15
	attribute [4]	1
texts [99]	text	единая россия вђ" википедия 18 1 2 в партия жуликов и воровв 18 1 3 события в сагре 18 2 утверждения о ru wikipedia org
	type	ExtLink
	rank	18.0
	attribute [0]	1
	attribute [1]	4932
	attribute [2]	17
	attribute [3]	0
	attribute [4]	1
texts [100]	text	матчасть
	type	ExtLink
	rank	1.0
	attribute [0]	1
	attribute [1]	0
	attribute [2]	0
	attribute [3]	0
	attribute [4]	2

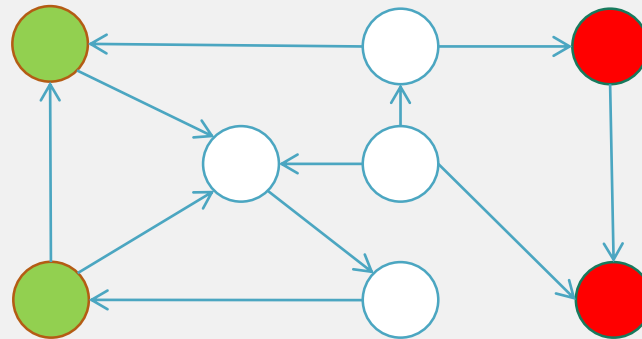


- Сколько вершин?
 - количество сайтов в интернете (50-100 млн)
 - количество документов в интернете (5-10 млрд)
- Сколько ребер?
 - ребер в 10 раз больше вершин
 - внутренних ссылок в 5 раз больше, чем внешних
- Как определить авторитетность страниц?
 - чем больше ссылок со страницы, тем лучше
 - чем больше ссылок на страницу, тем лучше

Простой итеративный алгоритм



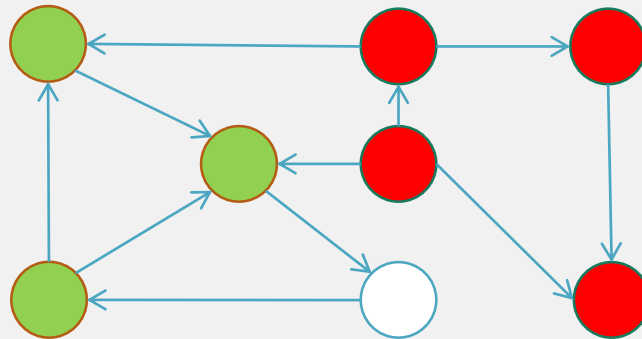
- Сайты: Хорошие, Плохие и Неизвестные
 - Если ссылаешься на Плохих – ты плохой
 - Если Хороший ссылается на тебя – ты хороший



Простой итеративный алгоритм



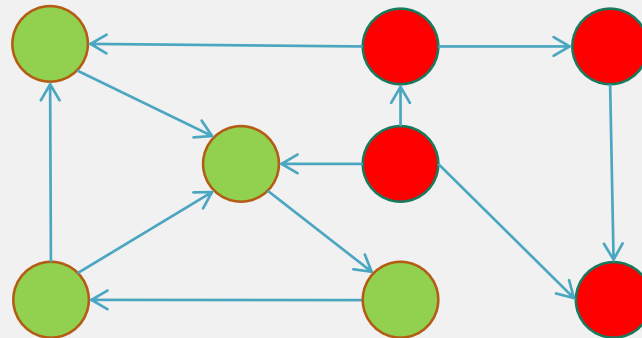
- Сайты: Хорошие, Плохие и Неизвестные
 - Если ссылаешься на Плохих – ты плохой
 - Если Хороший ссылается на тебя – ты хороший

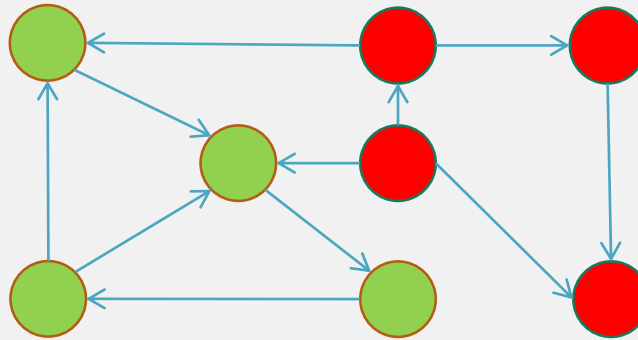


Простой итеративный алгоритм



- Сайты: Хорошие, Плохие и Неизвестные
 - Если ссылаешься на Плохих – ты плохой
 - Если Хороший ссылается на тебя – ты хороший





Нужна система на основе вероятностей

Hyperlink-Induced Topic Search (HITS)



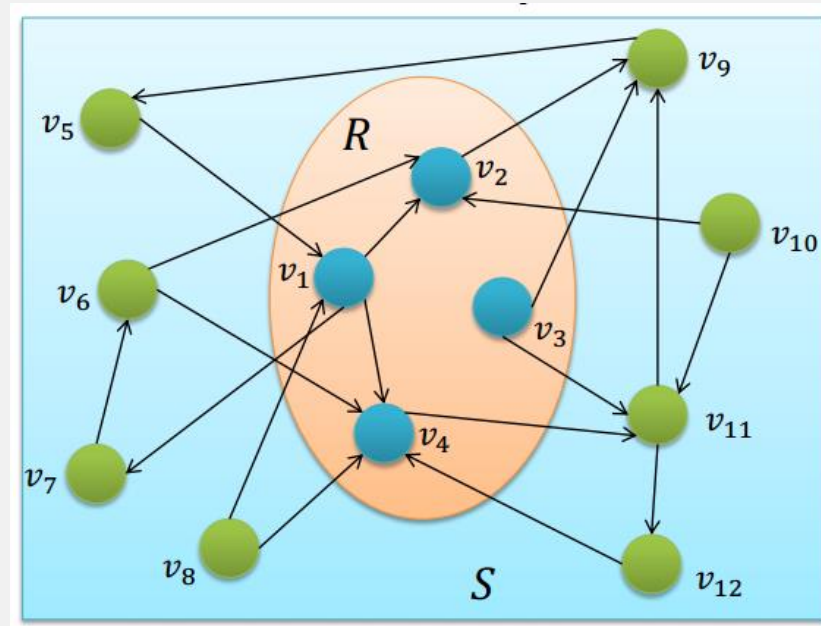
2 вида страниц

- Авторитеты – качественный контент
- Хабы – ссылки на авторитетов

2 гипотезы

- Хороший хаб ссылается на хороших авторитетов
- Хороший авторитет имеет ссылки с хороших хабов

Алгоритм HITS



Формируем 2 множества:

1. R – начальное множество узлов (например, результат поиска)
2. S – расширенное множество тех, кто ссылается на R

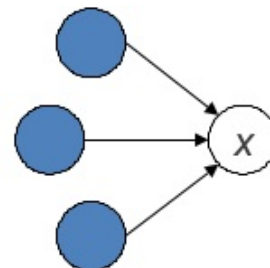
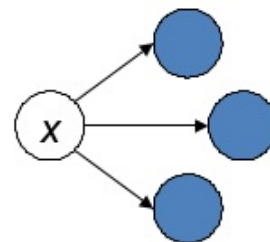


3. **Инициализируем:** $h(p) = 1$ $a(p) = 1$

4. Итерации:

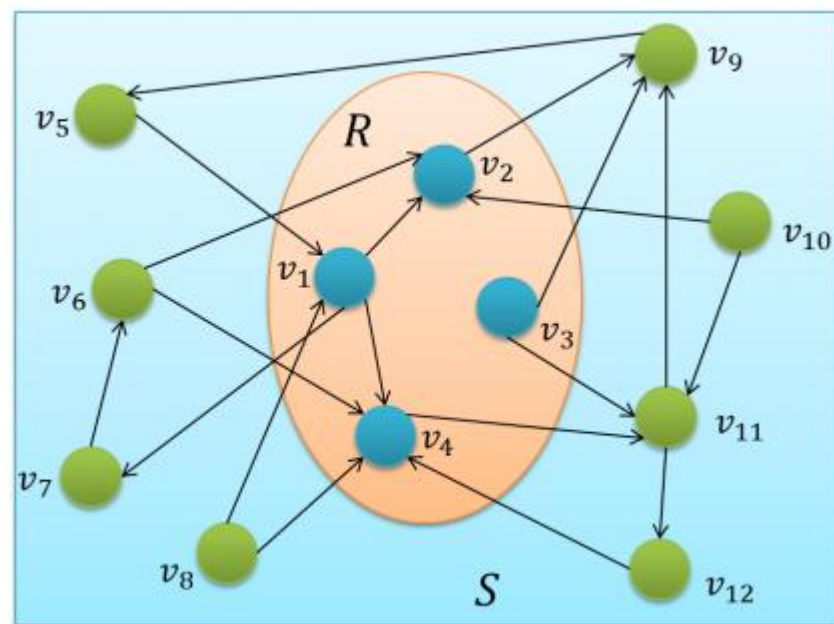
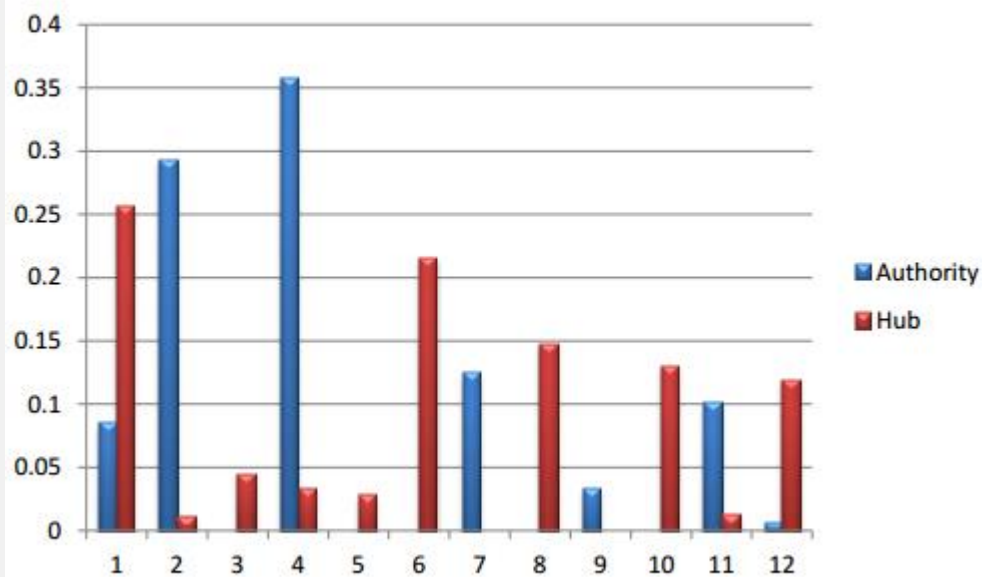
$$h(x) \leftarrow \sum_{x \rightarrow y} a(y)$$

$$a(x) \leftarrow \sum_{y \rightarrow x} h(y)$$



Достаточно 5 – 10 итераций

Алгоритм HITS



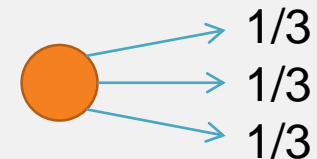
PageRank



- Модель случайного блуждателя:
 - Произвольная точка старта
 - На каждом шаге куда-нибудь переходим
 - С некоторой вероятностью

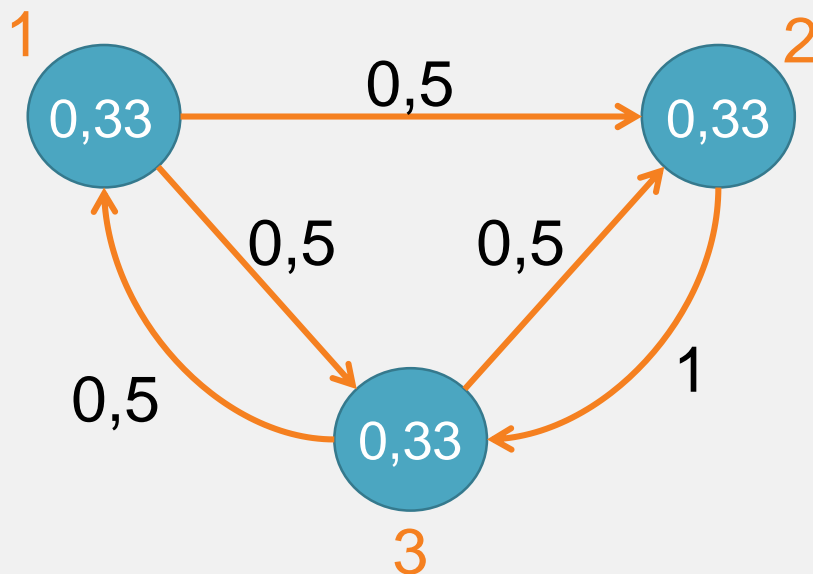


- С какой вероятностью странник
зайдет на наш сайт?



[Page et al, 1998]

PageRank: первый переход

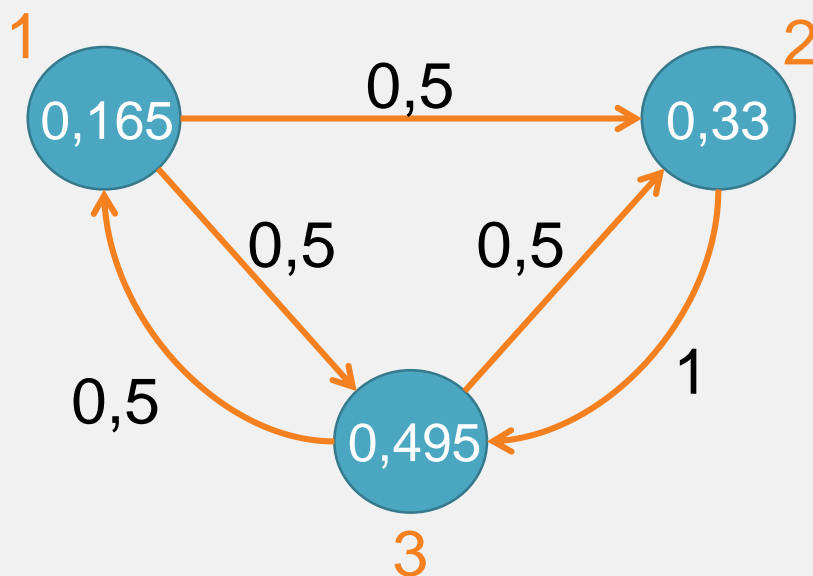


$$p_1 = 0,33 * 0,5 = 0,165$$

$$p_2 = 0,33 * 0,5 + 0,33 * 0,5 = 0,33$$

$$p_3 = 0,33 * 0,5 + 0,33 * 1 = 0,495$$

PageRank: второй переход

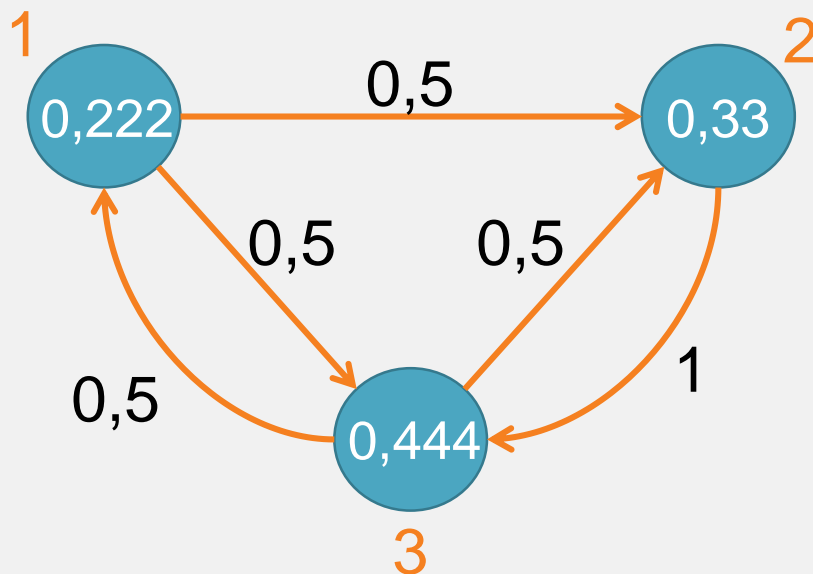


$$p_1 = 0,495 * 0,5 = 0,247$$

$$p_2 = 0,165 * 0,5 + 0,495 * 0,5 = 0,33$$

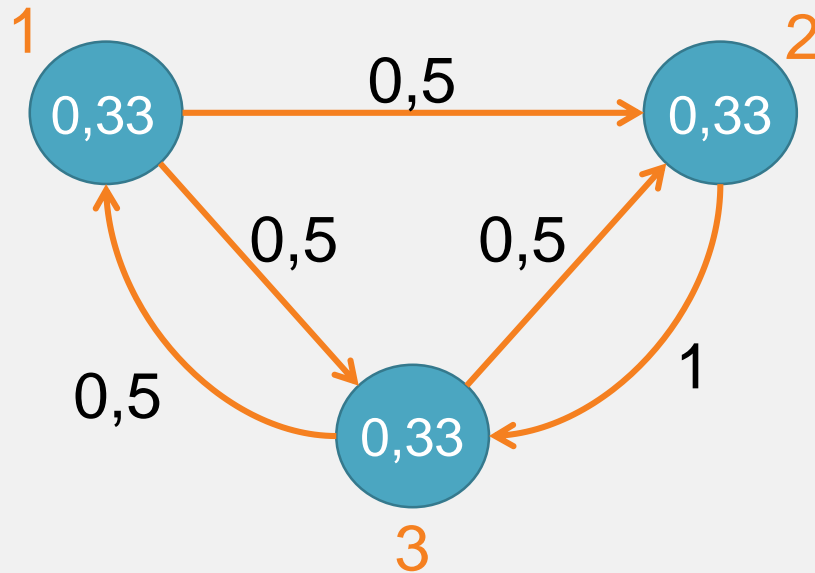
$$p_3 = 0,165 * 0,5 + 0,33 * 1 = 0,413$$

PageRank: 10-й переход



1	2	3	4	5	6	7	8	9	10
0,333333	0,166667	0,25	0,208333	0,229167	0,21875	0,223958	0,221354	0,222656	0,222005
0,333333	0,333333	0,333333	0,333333	0,333333	0,333333	0,333333	0,333333	0,333333	0,333333
0,333333	0,5	0,416667	0,458333	0,4375	0,447917	0,442708	0,445313	0,44401	0,444661

PageRank: переходим к матрицам



$$\begin{aligned} p_1 &= 0 \cdot 0,33 + 0 \cdot 0,33 + 0,5 \cdot 0,33 \\ p_2 &= 0,5 \cdot 0,33 + 0 \cdot 0,33 + 0,5 \cdot 0,33 \\ p_3 &= 0,5 \cdot 0,33 + 1 \cdot 0,33 + 0 \cdot 0,33 \end{aligned} = \begin{pmatrix} 0 & 0 & 0,5 \\ 0,5 & 0 & 0,5 \\ 0,5 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0,33 \\ 0,33 \\ 0,33 \end{pmatrix}$$

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

PageRank: определение



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

PageRank: $p = \lim_{k \rightarrow \infty} p^{(k)}$

$$p = P^T p$$

т. е. p — собственный вектор матрицы P^T



Цепь Маркова — последовательность случайных событий с конечным или счётным числом исходов, характеризующаяся тем, что при фиксированном настоящем будущее независимо от прошлого. Процесс в каждый момент времени находится в одном из состояний.

При этом, если он находится в состоянии с номером i , то он перейдет в состояние j с вероятностью p_{ij}

$P = \|p_{ij}\|$ - матрица переходов:

1. $p_{ij} \geq 0$
2. $\forall i \sum_j p_{ij} = 1$

Стационарное распределение в цепи Маркова



Эргодическое (стационарное) распределение -
распределение $\alpha = (\alpha_1 \dots \alpha_n)$, такое что $\alpha_i > 0$ и
 $\alpha_j = \lim_{n \rightarrow \infty} p_{ij}^{(n)}$

Эргодическая теорема: в сильно связной и апериодической цепи Маркова с матрицей переходов P существует единственное стационарное распределение p :

$$p = Ap, A = P^T$$

PageRank: висячие узлы (dangling nodes)

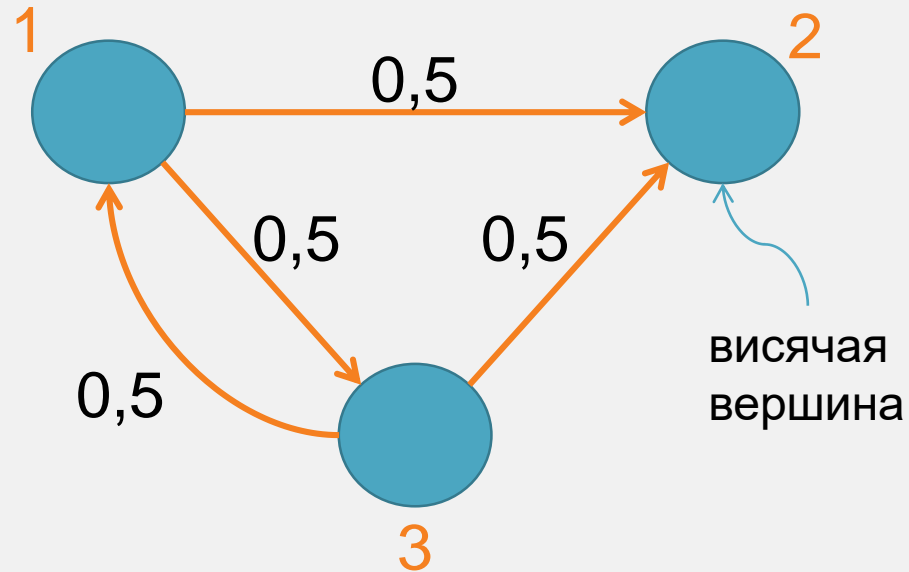


Причины:

1. На странице нет ссылок
2. Страницу не обкачивали

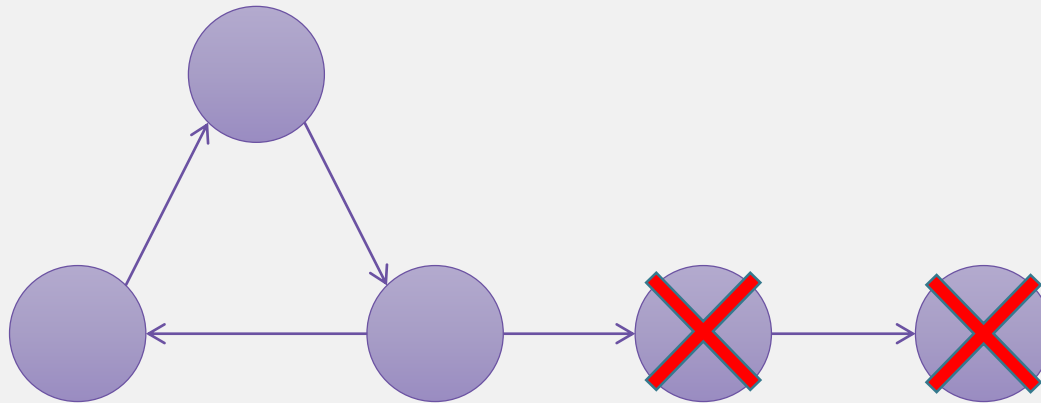
В чем проблема?

$$\sum_j p_{ij} = 0, \text{ если } i - \text{висячая}$$



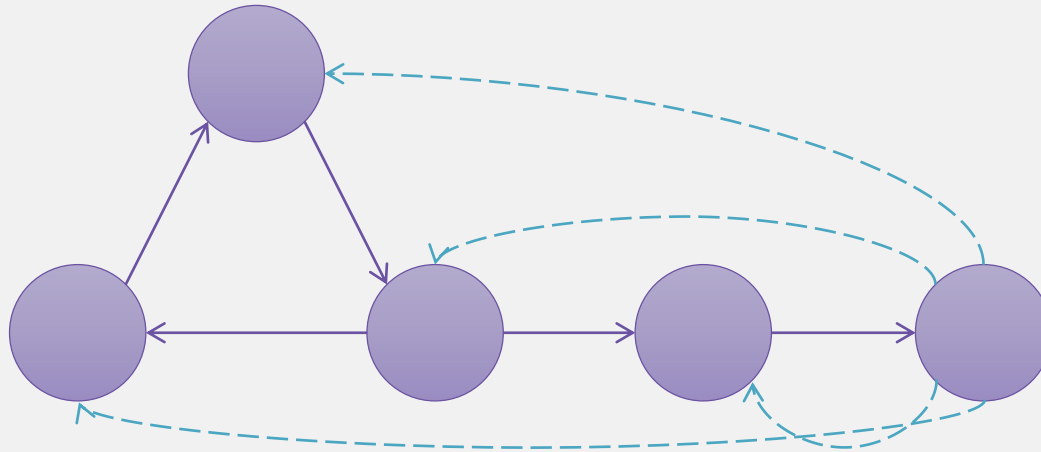
В реальном индексе до 60% висячих вершин

Висячие узлы: метод удаления



- Т.к. вершины не оказывают влияния на остальной граф, то удаляем их
- Удаление висячих вершин порождает новые, поэтому нужно несколько итераций
- После расчета PageRank эти вершины нужно вернуть

Висячие вершины: связь со всеми другими



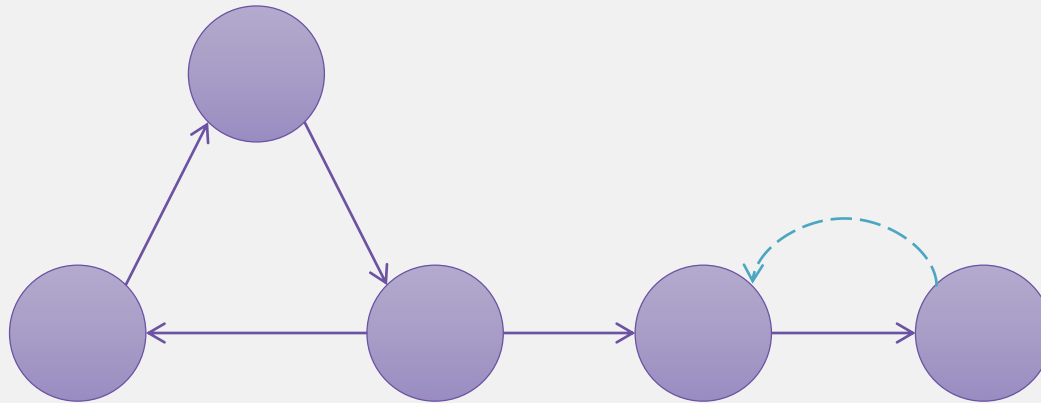
- Попав на страницу без ссылок, пользователь перескакивает на любую другую страницу сети случайным образом
- Вероятность перехода может быть одинаковой, а могут быть разными

$$P' = P + d \cdot v^T$$

$d[i]=1$ если i – висячий узел, 0 – иначе

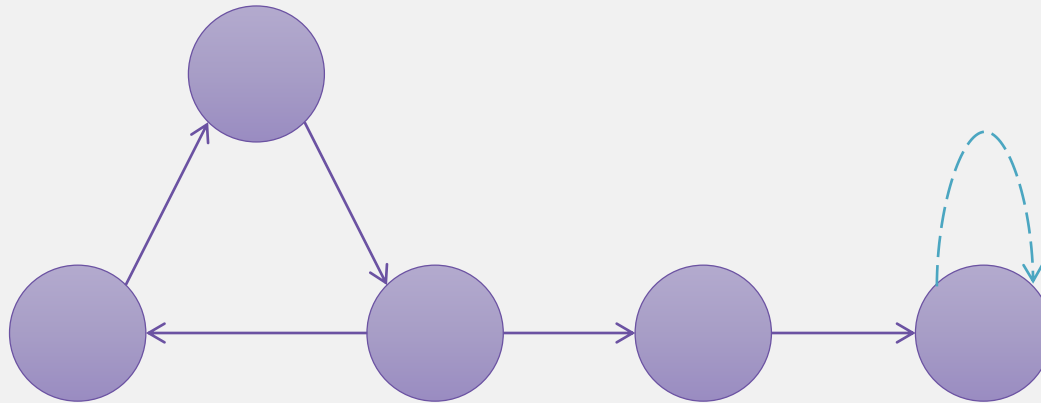
$v[j]$ – вероятность перехода на страницу j с висячего узла

Висячие вершины: шаг назад



- Соответствует нажатию кнопки “назад” в браузере
- Поощряет страницы с большим числом ссылок на висячие узлы

Висячие вершины: Петля



- Матрица приходит в порядок
- Но висячие узлы собирают много PageRank
- Нужно скорректировать результаты:

[Scaling Personalized Web Search](#)



- Пользователь переходит по ссылками с вероятностью d , а с вероятностью $(1-d)v[j]$ переходит на страницу j
- $d \sim 0.85-0.9$
 - чем больше d , тем точнее результаты, но медленнее сходимость
 - чем меньше d , тем чувствительней к спаму
- Демократический вектор телепортации: одинаковые значения для всех страниц
- Аристократический: выше вероятность у хороших сайтов
- $v[j] \neq 0$
 $P'' = dP' + (1 - d)E, E = (1 \dots 1) \cdot v^T, 0 < d < 1$

Ускорение метода степеней: Метод Гаусса-Зейделя (Gauss-Seidel)



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

- Вместо $p^{(k)}$ используем уже посчитанные $p^{(k+1)}$
- Уменьшение числа итераций на 40%
- Сложно применять в параллельных вычислениях

Ускорение метода степеней: Метод экстраполяции



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

- После нескольких итераций строим аппроксимацию и используем ее вместо $p^{(k)}$
- Уменьшение числа итераций на 30%

Ускорение метода степеней: Адаптивный метод



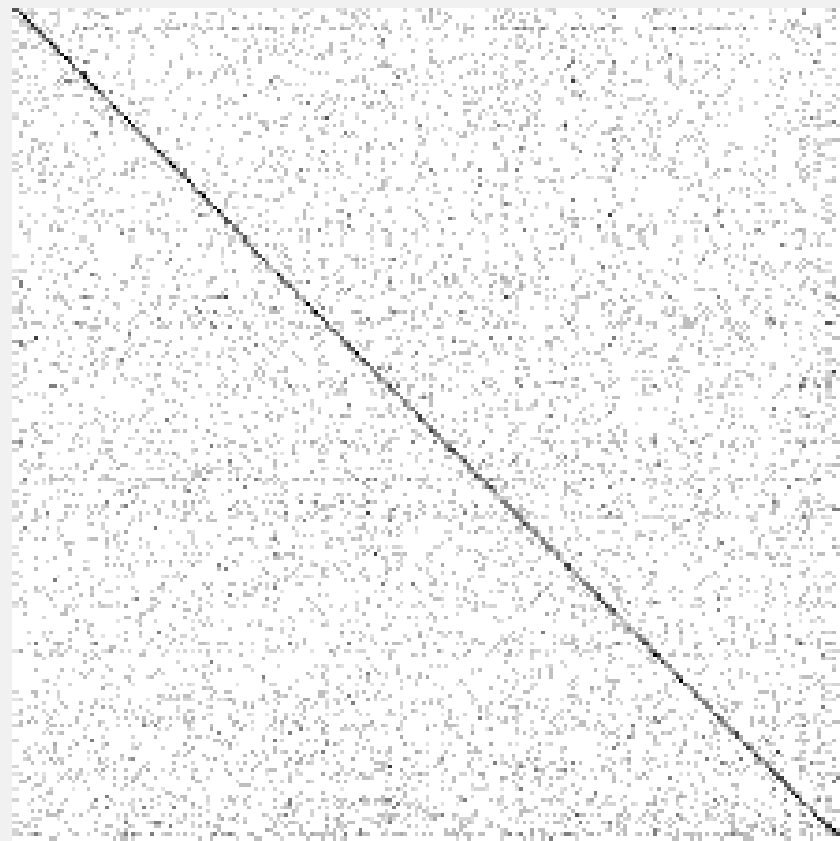
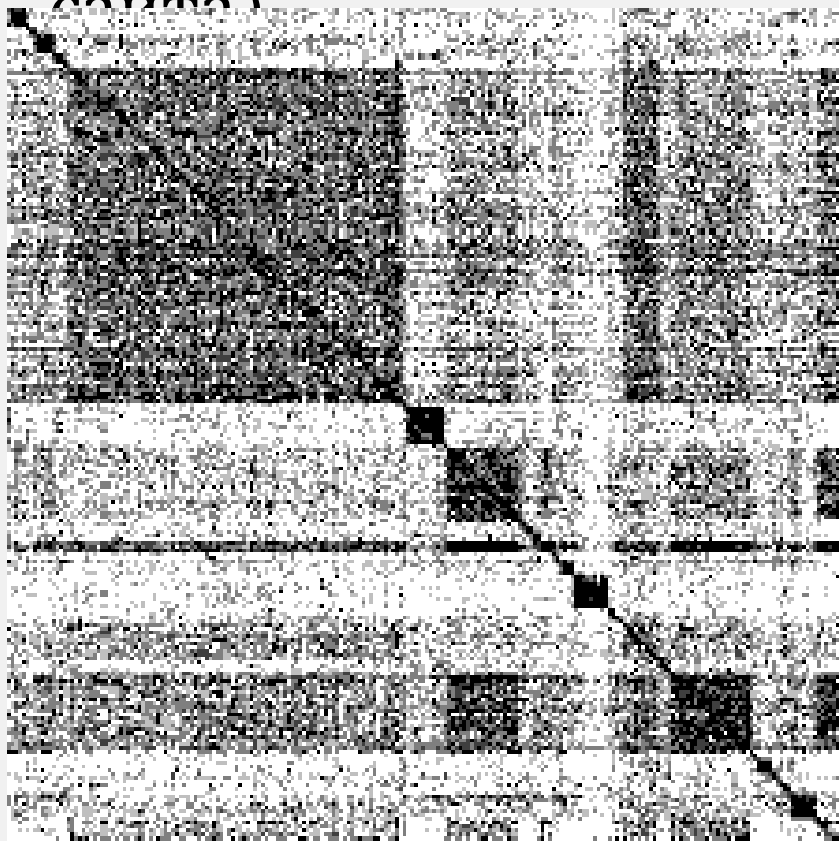
$$p^{(k+1)} = P^T p^{(k)}$$

- Некоторые элементы p сходятся значительно быстрее остальных
- Если $p^{(k+1)}[i] - p^k[i] < \varepsilon$, то значение для i найдено
- Ускорение расчетов на 20%

Ускорение метода степеней: Метод блочной структуры



- ID страницы: (id сайта; id страницы внутри сайта)

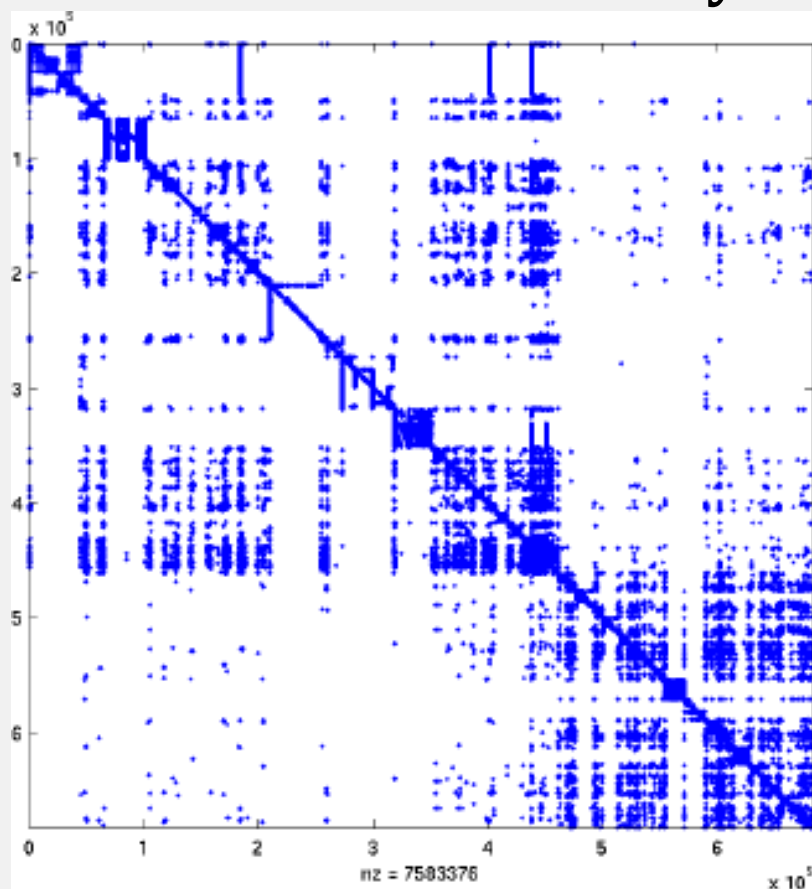


*.com

Ускорение метода степеней: Метод блочной структуры



stanford.edu и berkeley.edu:



Более 90% ссылок в интернете
являются внутресайтовыми

Exploiting the Block Structure of the Web for Computing PageRank (Sepandar D. Kamvar и др.)

Ускорение метода степеней: Метод блочной структуры



1. Строим граф сайтов. Считаем PageRank для него.
2. Считаем PageRank по внутренним ссылкам для каждого сайта
3. Локальный PageRank умножается на рейтинг сайта и записывается в вектор Z
4. Вектор Z используется как начальный для классического PageRank

Ускорение метода степеней: Метод блочной структуры



1. Локальный PageRank можно вычислять в памяти
2. Вычисление локального PageRank можно распараллелить
3. Алгоритм для графа страниц сходится достаточно быстро

Общее число итераций сокращается на 50%!



- PageRank по графу сайтов можно использовать не только для вычисления обычного PageRank
- Можно улучшать ранжирование:
 - **проблема:** PageRank поощряет старые страницы и недооценивает новые
 - **решение:** даем бонус за хороший рейтинг сайта
- SiteRank более устойчив к поисковому спаму



$$p = Ap, A = P^T$$

- p – собственный вектор матрицы A
- Может быть найден путем решения системы уравнений
- Правда система уравнений огромная(;

PageRank: параллелизация



1. Выделяются сильно связанные компоненты. PageRank на каждой вычисляется параллельно
2. Используется блочная структура графа
3. MapReduce (см. лекцию №6 в курсе про Hadoop)

PageRank: эволюция графа



- Как не пересчитывать весь PageRank, если меняется только часть графа?
- В “Incremental PageRank Computation on Evolving Graphs” Prassana Desikan предлагает способ:
 1. Выделить неизмененную часть графа
 2. Удалить эту часть, оставив только граничные узлы
 3. Посчитать PageRank оставшейся части
 4. Отмасштабировать результат на весь граф
- Результат получается неточным
- Каждую неделю меняется до 25% ссылок, поэтому лучше пересчитывать весь граф



- Пусть дана категория сайтов:
 - культура, спорт, политика и т.д
- T – множество страниц из этой категории
- Делаем вектор телепортации $v[i]$:
 - $v[i]$ очень мало, если i не в T и большое в ином случае

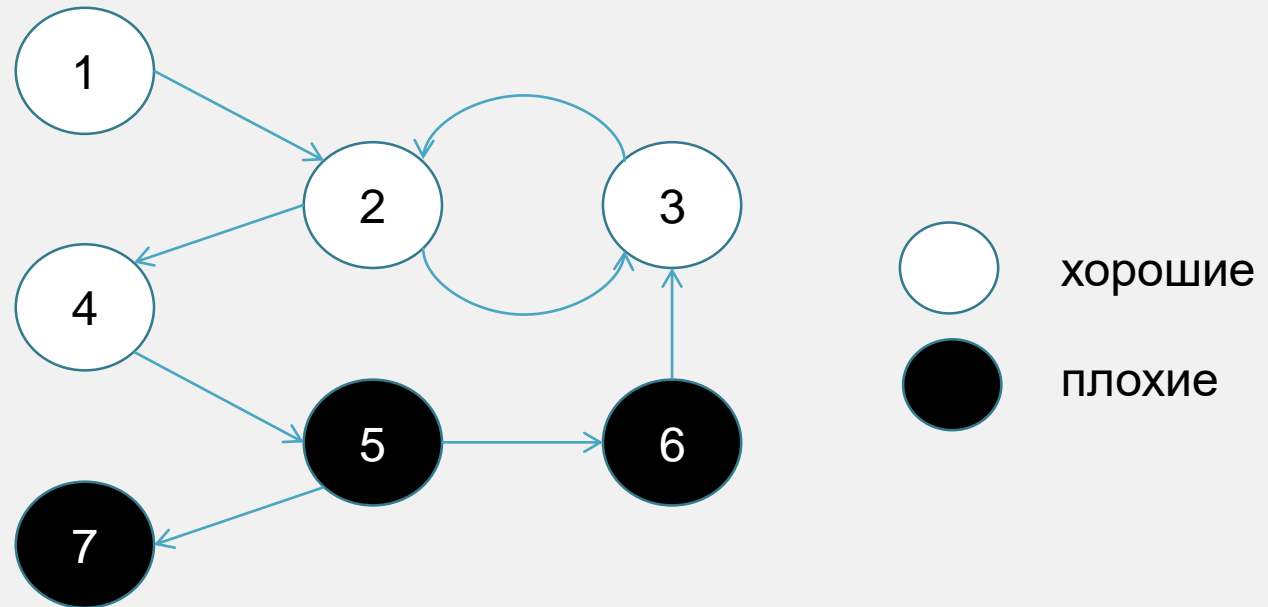
$$P'' = dP' + (1 - d)E, E = (1 \dots 1) \cdot v^T, 0 < d < 1$$

PageRank: персонализация



- Пусть есть n категорий сайтов (~ 100)
- Рассчитаем для каждой тематический PageRank
- Если вектор k отражает заинтересованность каждого пользователя к разным тематикам, то PageRank для него:

$$PageRank_{person} = k[0] * p_0 + k[1] * p_1 + \dots + k[n] * p_n$$

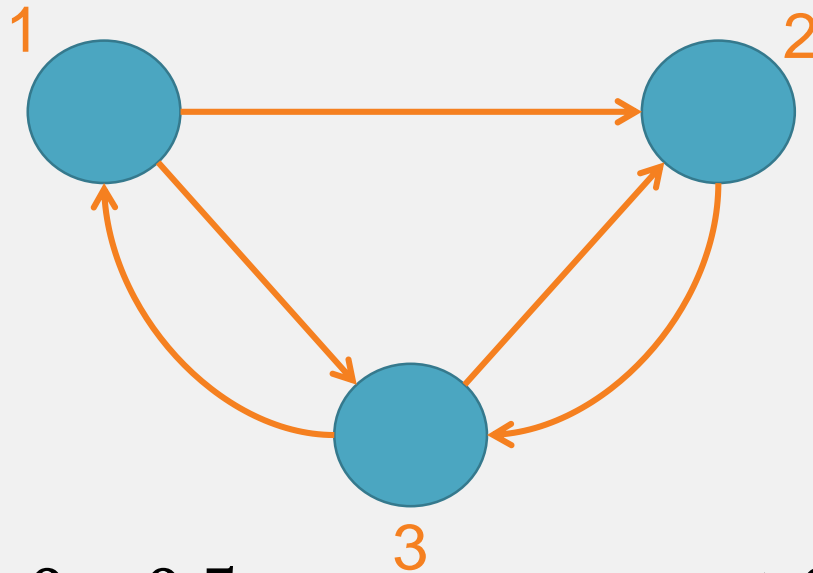


- Разметить все страницы трудно
- Можно разметить лишь ограниченное множество
- Дальше смотрим, как сигнал от хороших сайтов распространяется по графу



- Популярные сайты (топ 1000 в каталогах)
 - низкая полнота
- Результат работы поисковиков
 - низкая точность
- Результат PageRank
 - подвержен спаму
- Обратный PageRank

Обратный PageRank

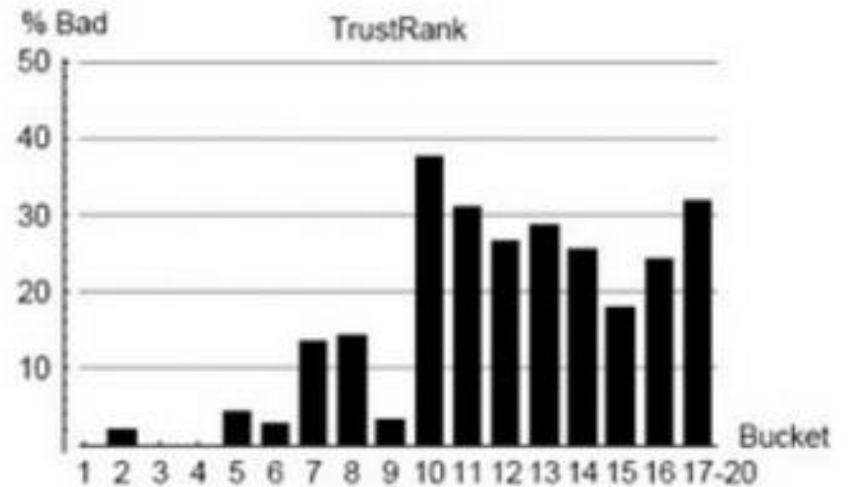
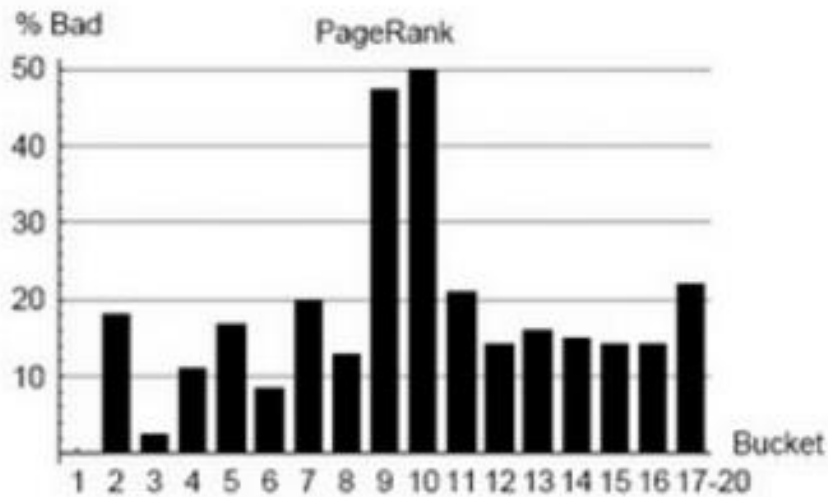


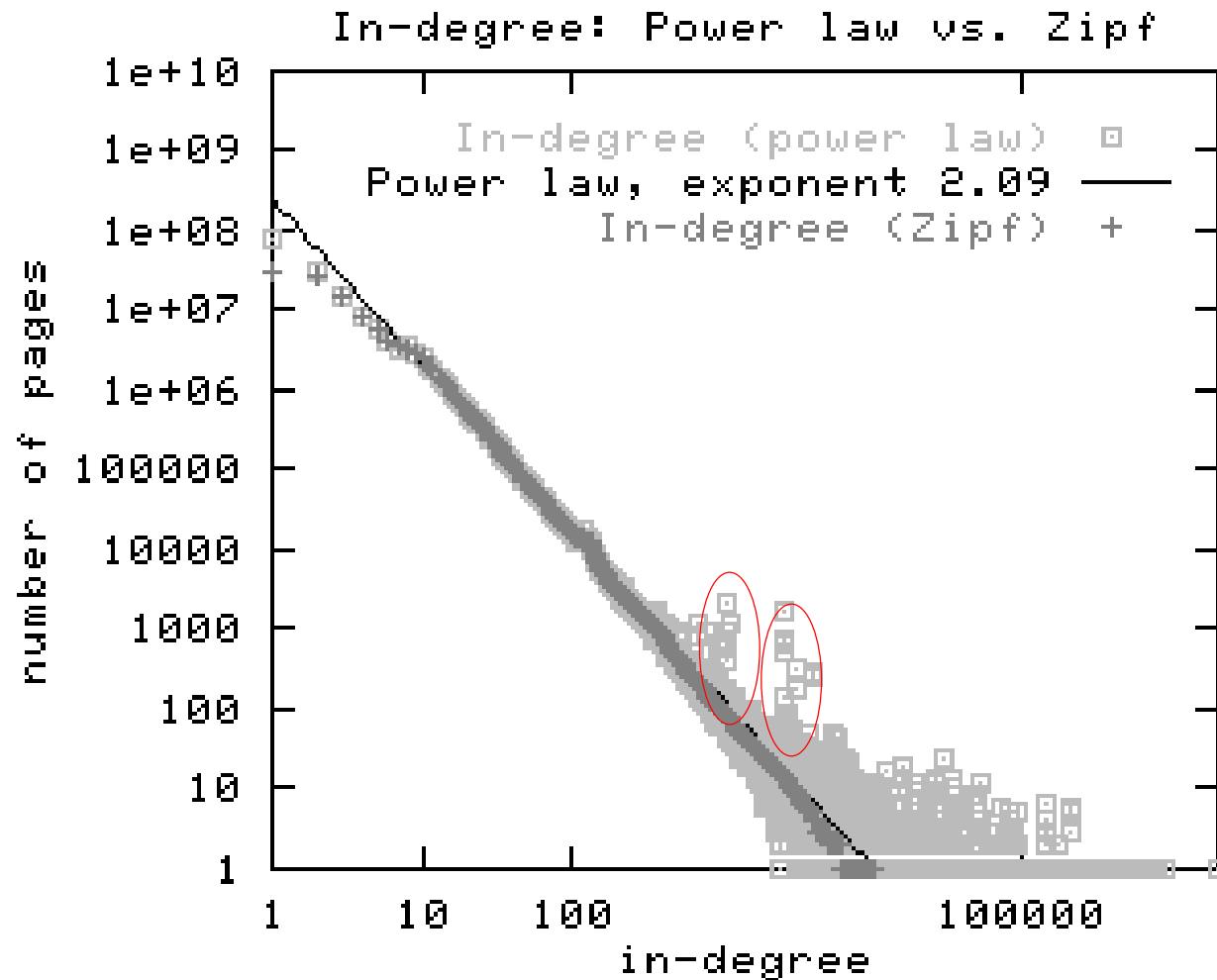
$$P = \begin{pmatrix} 0 & 0 & 0,5 \\ 0,5 & 0 & 0,5 \\ 0,5 & 1 & 0 \end{pmatrix}$$

$$U = \begin{pmatrix} 0 & 0,5 & 0,5 \\ 0 & 0 & 0,5 \\ 0,5 & 1 & 0 \end{pmatrix}$$

$$p_r^{(k+1)} = U^T p_r^{(k)}$$

PageRank vs TrustRank







Multi Search university [Next! \[national parks\]](#)

10 results

clustering on

Search

Query: **university**
11 Results Returned
Showing Results From 0 to 10

Stanford University Homepage
74.79% <http://www.stanford.edu/>
4K - 2591993 - 010397

Stanford University: Portfolio Collection
65.78% <http://www.stanford.edu/home/administration/portfolio.html>
3K - 2591993 - 010397

University of Illinois at Urbana-Champaign
73.26% <http://www.uiuc.edu/>
15K - 1313096 - 010397

Indiana University
68.38% <http://www.indiana.edu/>
1K - 0923996 - 010397

University of California, Irvine
68.07% <http://www.uci.edu/>
3K - 1313096 - 010397

University of Minnesota
67.05% <http://www.umn.edu/>
0K - 1316996 - 010397

Iowa State University Homepage
66.66% <http://www.iastate.edu/>
3K - 1313096 - 010397

The University of Michigan
66.35% <http://www.umich.edu/>
1K - 2591993 - 010397

Mississippi State University
66.35% <http://www.msstate.edu/>
3K - 2591993 - 010397

Northwestern University: NUInfo
66.15% <http://www.nwu.edu/>
3K - 1314996 - 010397

next 10

Optical Physics at the University of Oregon
Oregon Center for Optics in Science and Technology. Department of Physics, University of Oregon, Eugene OR 97403. Research Groups: Carmichael Group....
<http://opci.b.uoregon.edu/> - size 1K - 16 Dec 96

Carnegie Mellon University - Campus Networking
Departments. Data Communications. Data Communications is responsible for installing and maintaining all on campus networking equipment and all of...
<http://www.net.cmu.edu/> - size 4K - 19 Aug 95

Wesleyan University Computer Science Group Home Page
Computer Science Group. Wesleyan University. Welcome to the home page of the Computer Science Group at Wesleyan University. We are administratively within.
<http://www.cs.wesleyan.edu/> - size 3K - 15 Apr 96

Keio University Shonan Fujisawa Campus (SFC)
B\$3\$N%ZIEFnF#Bt%-%c%e%Q%99 (B(SFC) \$B\$N (BWWW \$B% \$BCmOU=q\$- (B \$B\$FI\$s\$G\$/\$@5\$ \$!# (B. Nihongo | English. SFC \$B>pJs (B. [\$B%e%G%#%*%/%e%? !*...
<http://www.sfc.keio.ac.jp/> - size 3K - 5 Feb 97

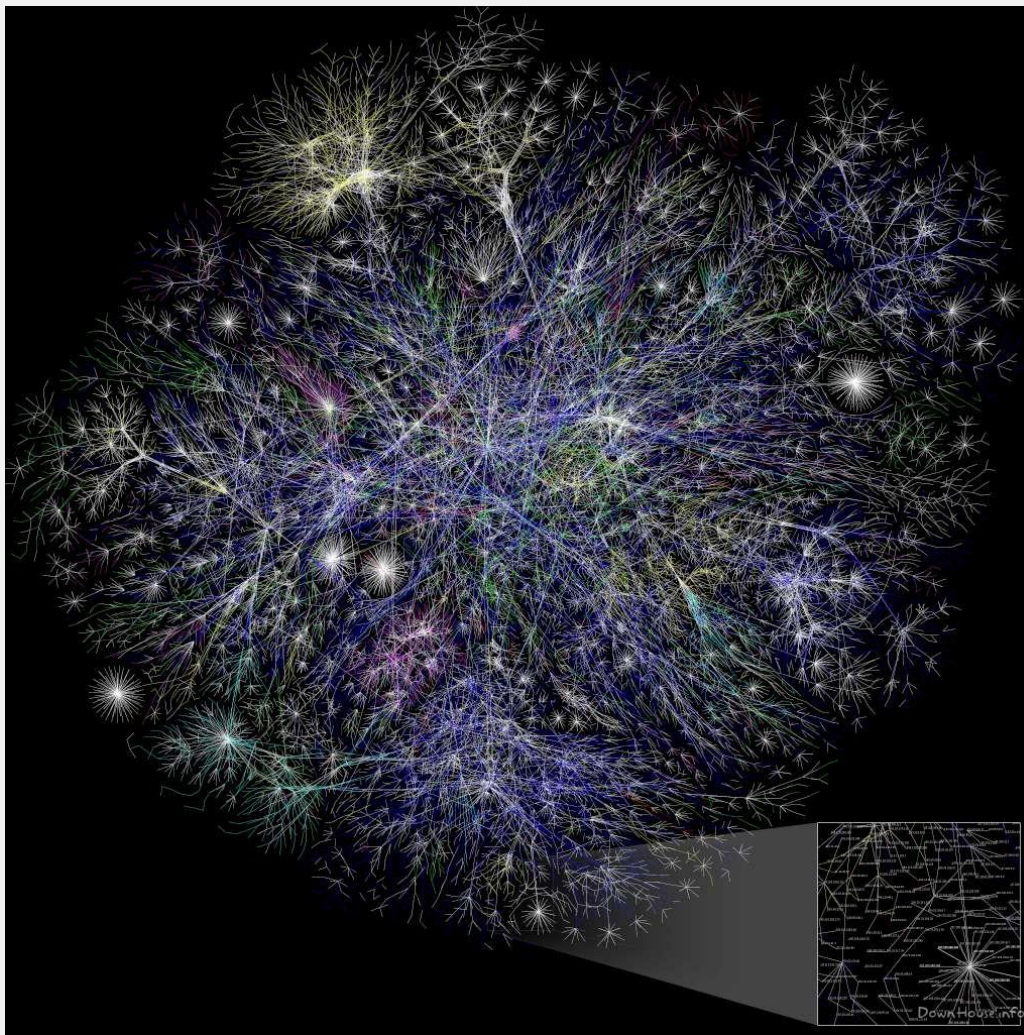
School of Chemistry, University of Sydney
The School of Chemistry. School of Chemistry, University of Sydney, NSW 2006 Australia International Phone: +61-2-9351-4504 Fax: +61-2-9351-3329 Australia.
<http://www.chem.su.oz.au/> - size 4K - 25 Feb 97

Mankato State University
The Campus Athletics, Campus Tour, Bookstore, Maps, Current Events... Admission & Registration Admissions, Financial Aid, Registrar's, Graduate...
<http://www.mankato.msus.edu/> - size 3K - 27 Nov 96

St. Ambrose University
Main Index: Academic Departments. Administrative Services. Campus News. Computing Services. Galvin Fine Arts Center. Internet Connections. Library...
<http://www.sau.edu/> - size 3K - 4 Feb 97

University of Washington ECSEL Projects

Ссылочный граф



Отмечайтесь и оставляйте отзыв

**Спасибо за
внимание!**

Евгений Чернов

e.chernov@corp.mail.ru

Домашнее задание



Реализовать алгоритмы HITS и PageRank с помощью MapReduce для сайта lenta.ru

Исходные данные: html страницы с lenta.ru ([скачать](#))

Формат данных:

- id документа + base64(gzip(html))
- urls.txt содержит id страниц

На выходе:

1. Граф ссылок
2. ТОП30 по алгоритму HITS и PageRank (5 баллов)
3. MapReduce код для расчета авторитетности страниц с помощью этих алгоритмов (10 баллов)

Адрес: sfera.linkgraph@mail.ru

Срок: до 12 апреля



Требуется: Реализовать алгоритм HITS для сериалов с afisha.mail.ru

Исходный данные: Граф ссылок: [скачать](#)

На выходе: Самый авторитетный актер и самый “каталожный” сериал (в терминах HITS)