



ТЕХНОСФЕРА

Особенности web-поиска. Спайдер.

Сергукова Юлия,
программист отдела инфраструктуры проекта
Поиск@Mail.Ru

План лекции:

1. Web-поиск

- 1. Историческая справка**
- 2. Немного о рекламе**
- 3. Схемы**

2. Поисковый спайдер

- 1. Постановка задачи**
- 2. Выкачка**
- 3. Обновление**
- 4. Хранение**

Не забывайте о практике

Интернет и WWW

Интернет != WWW

Интернет и WWW

интернет != WWW

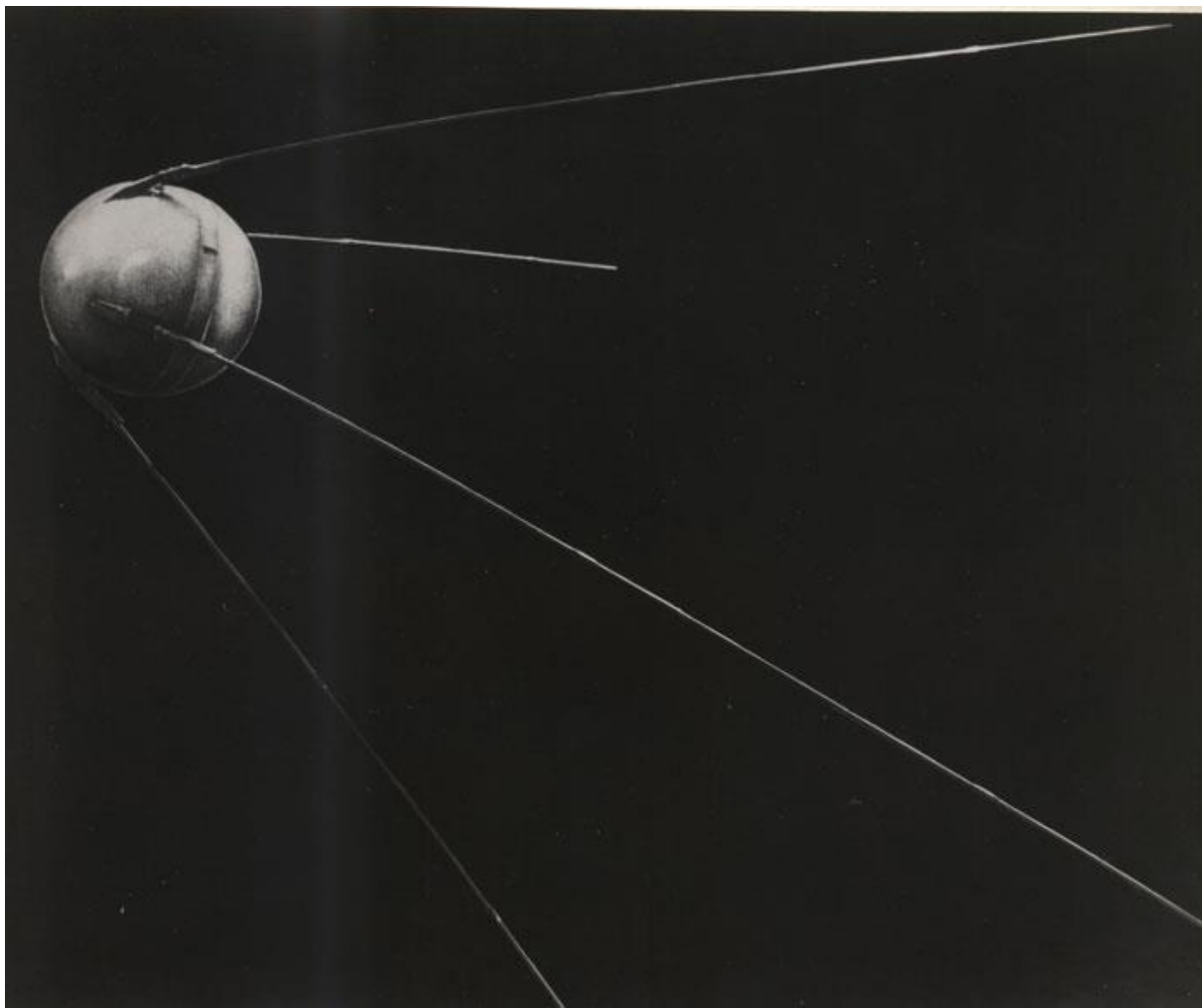
Разные уровни абстракции:

интернет - система объединенных компьютерных сетей

www - распределенная система, предоставляющая доступ к документам, расположенным на разных устройствах, подключенных к интернету

С чего всё началось?

С чего всё началось?



С чего всё началось?

29 октября 1969г. - рождение интернета (Калифорния - Массачусетс)

С чего всё началось?

1981г. - 200 машин. Интернет + www. IBM выпускает первый массовый (!) компьютер



С чего всё началось?

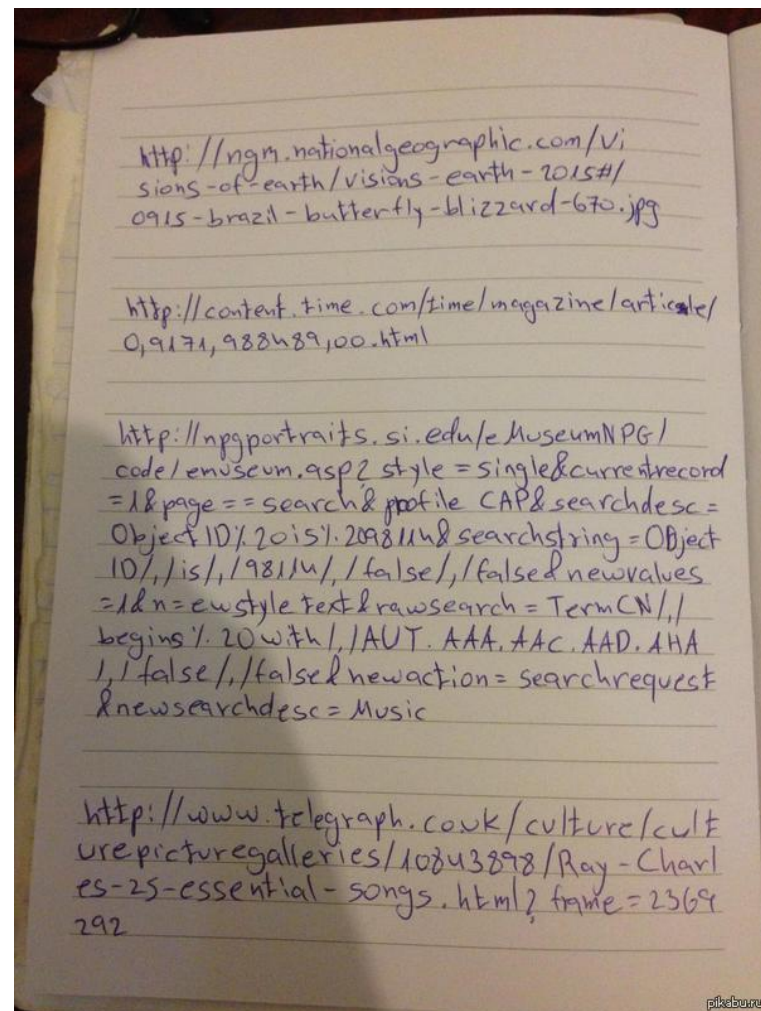
15 марта 1985г. - первый зарегистрированный домен: symbolics.com (в 2009г. перепродали - теперь рекламный сайт)

6 августа 1991г. - первый сайт. До сих пор функционирует:

<http://info.cern.ch/hypertext/WWW/TheProject.html>

История ПС

Данных мало - на странице университета NCSA (Иллинойс) поддерживается список "What's new". Вручную.



Поисковые системы

1990г. - первая поисковая программа Archie. Поиск по заголовкам файлов на FTP-серверах.

Поисковые системы

1993г.

AliWeb (Archie-like indexing for the WEB) - готовые индексы от администрации сайтов. Первая ПС

W3Catalog - не обкачивает сайты, а использует чужие списки страниц. Первый агрегатор

WWW (WorldWideWebWanderer) – первый поисковый робот. Цель - узнать все известные страницы.

Поисковые системы

1994 – keyword-based системы (не полнотекстовый поиск, но по множеству слов, связанных с документом)

Keyword-based системы

1. "Найди мне то, что я сказал" (сейчас "найди мне то, что я хотел")

Keyword-based системы

1. "Найди мне то, что я сказал" (сейчас "найди мне то, что я хотел")
2. Не последняя роль - содержимое тега meta-keywords

```
<head>  
<meta charset="UTF-8">  
<meta name="description" content="Free Web tutorials">  
<meta name="keywords" content="HTML,CSS,XML,JavaScript">  
<meta name="author" content="Hege Refsnes">  
</head>
```

КТО-ТО ИХ ДО СИХ ПОР ИСПОЛЬЗУЕТ:

<http://www.ultersuite.ru/articles/meta/>

а Google нет:

<https://webmasters.googleblog.com/2009/09/google-does-not-use-keywords-meta-tag.html>

Поисковые системы

1994 – keyword-based системы

AltaVista (до 8.07.2013) - первый короткий домен, первая "легкая" заглавная страница

Excite (<http://www.excite.com/>)

InfoSeek (в 1998 куплен The Walt Disney Company)

InkToMi (23.12.2002 поглощен Yahoo.com)

Поисковые системы

Люди начинают искать не конкретные файлы, а конкретную информацию.

Многообразие запросов.

Поисковые системы коммерциализируются.

Поисковые системы

1996 - “sponsored search”

Поисковые системы

1996 - “sponsored search”

место в выдаче по конкретному ключевику зависит от того, сколько вы за него заплатили

чем популярнее слово, тем дороже на нем продвигаться: **casino** было очень дорогим

Поисковые системы

1998г - GoTo

8.11.2001 - переименован в Overture.com;
поглотил AltaVista

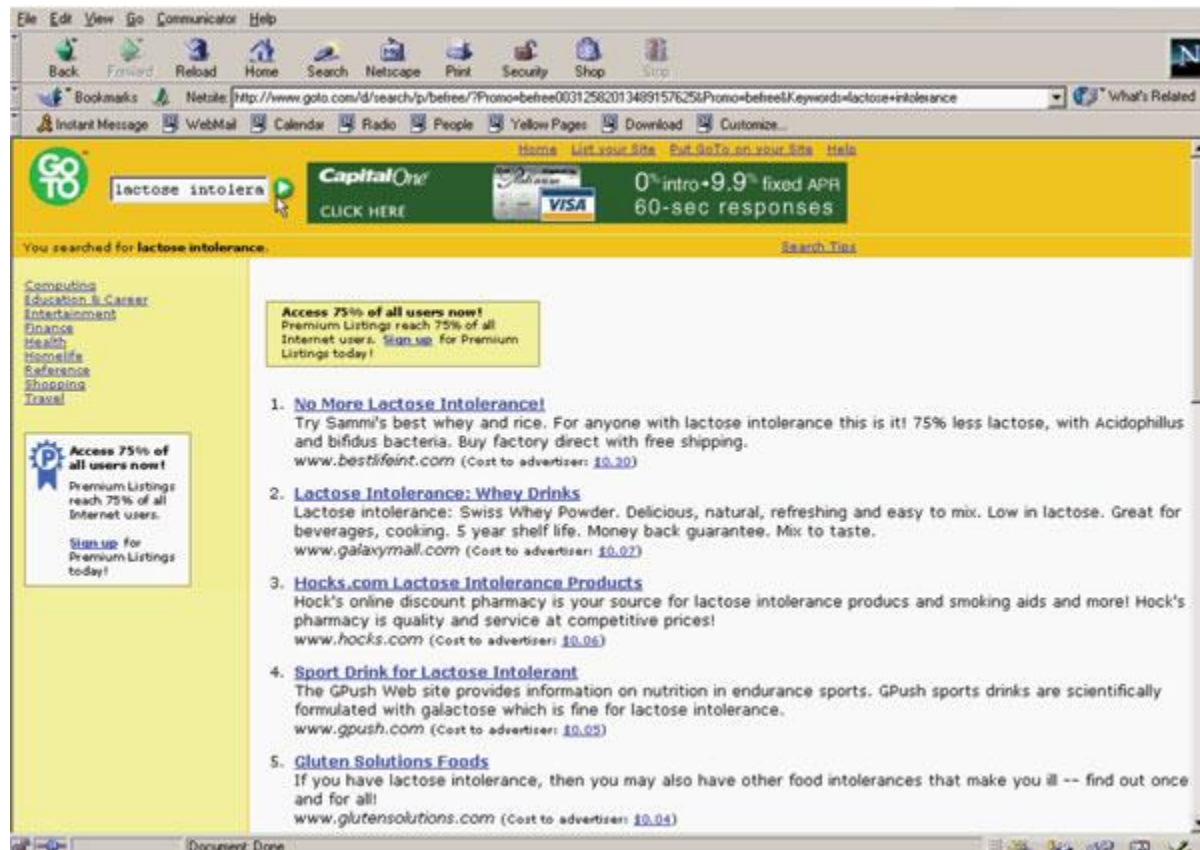
7.10.2013 - поглощен Yahoo.com

Sponsored search

Платят не только за то, **что** показывать, но и за то, **где** и **как**:

1. Платят только за переходы пользователя с ПС на сайт
2. Кто больше платит, тот выше в выдаче (выгода для ПС)
3. Чем выше в выдаче сайт, тем больше вероятность перехода (выгода для владельцев сайтов)

Начало эры поисковой рекламы



Google и PageRank

1998г.

Результаты выдачи ранжируются по своему "качеству" и релевантности запросу.

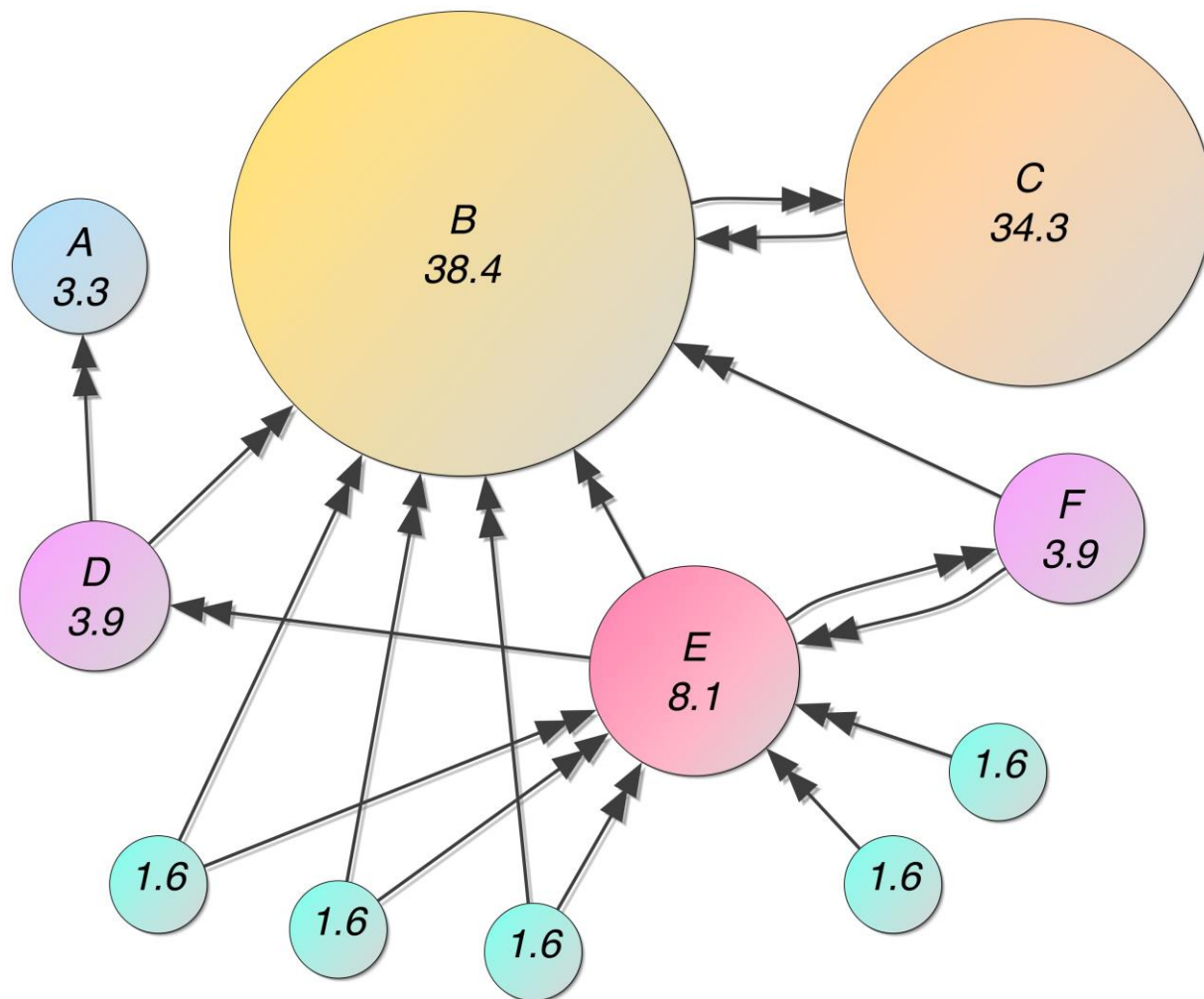
Релевантность - соответствие ожиданиям пользователя

Google и PageRank

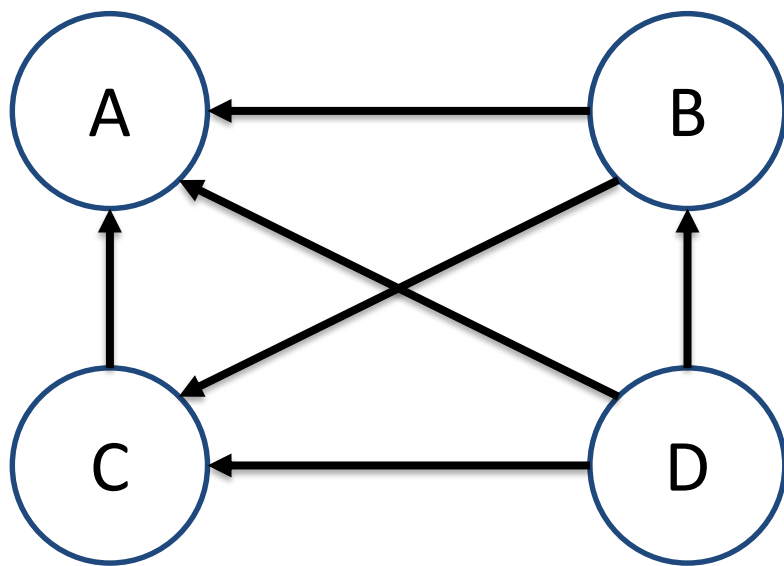
1. Содержание страниц (не только keywords)
2. Популярность страниц (индекс цитирования)
 - Linkfarm – фермы по "разведению" ссылок
 - Rank-sculpting – искусственное наращивание ранка (nofollow)

PageRank

PageRank

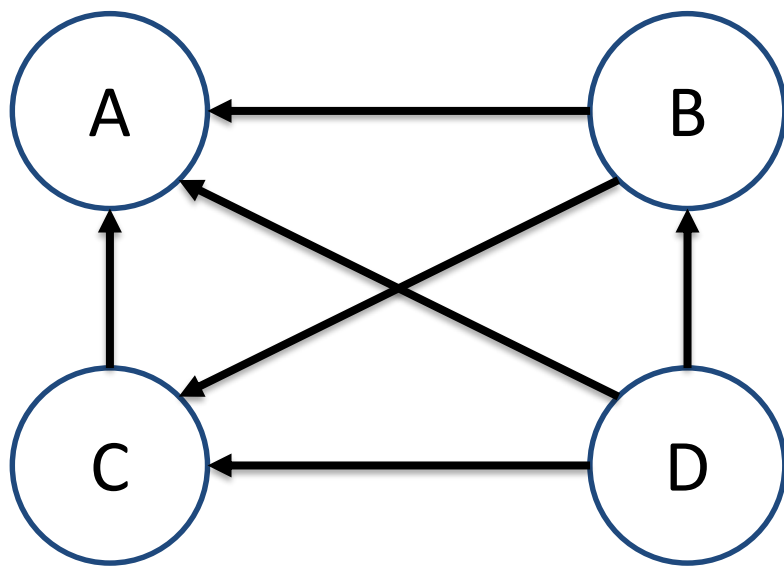


PageRank



PageRank

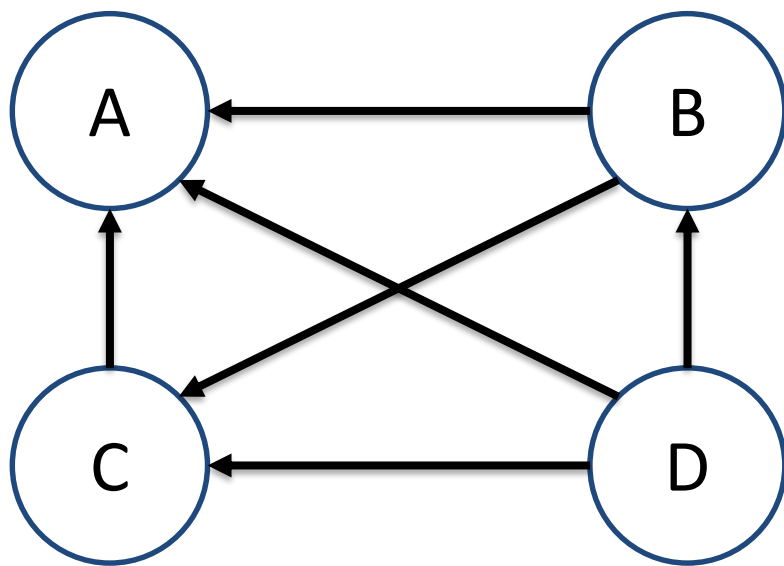
Стартовое значение: $PR(P) = 1/N$



PageRank

Чем больше ссылаются, тем лучше:

$$PR(A) = PR(B) + PR(C) + PR(D)$$

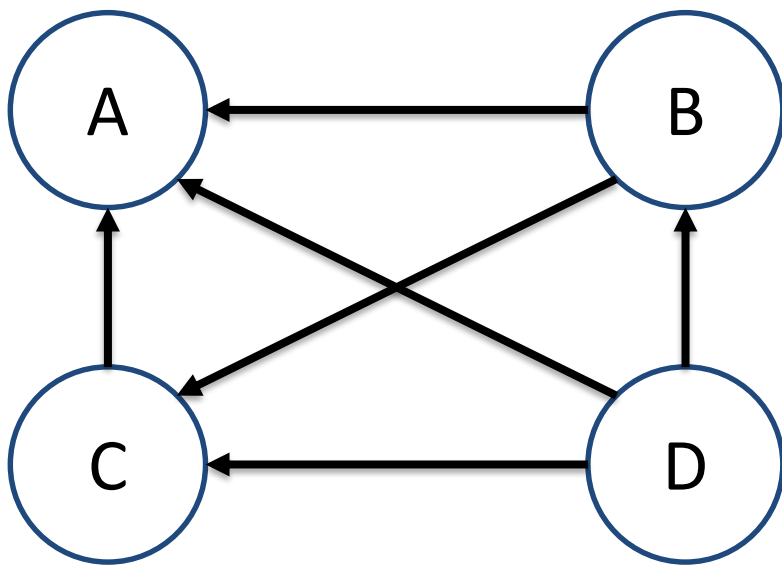


PageRank

Чем больше ссылаются, тем лучше:

$$PR(A) = PR(B) + PR(C) + PR(D)$$

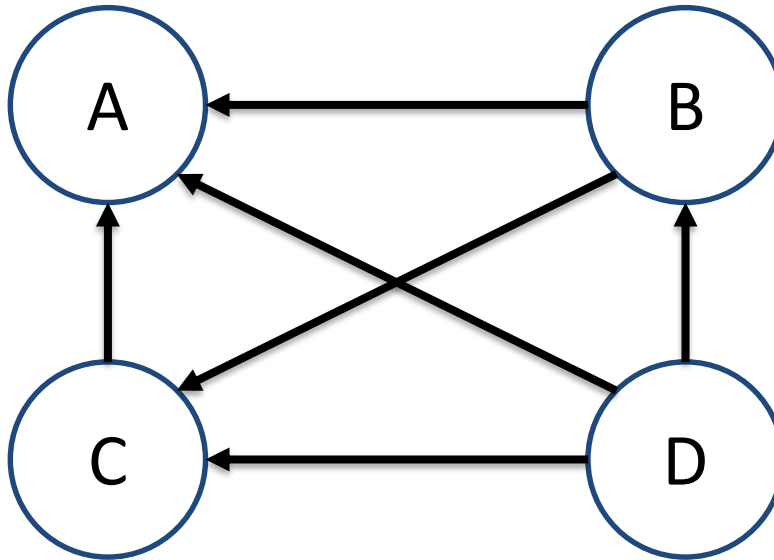
Но некоторые ссылки важнее: С или D?



PageRank

Чем больше ссылаются, тем лучше:

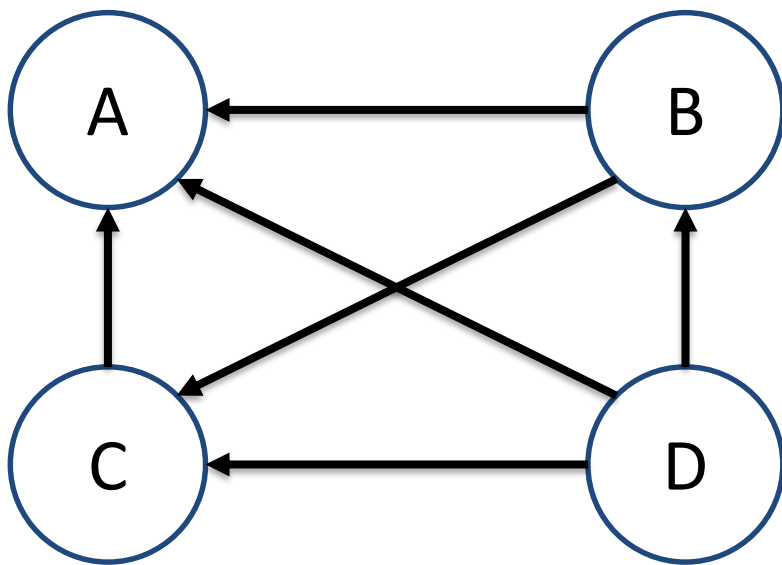
$$PR(A) = \frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)}.$$



PageRank

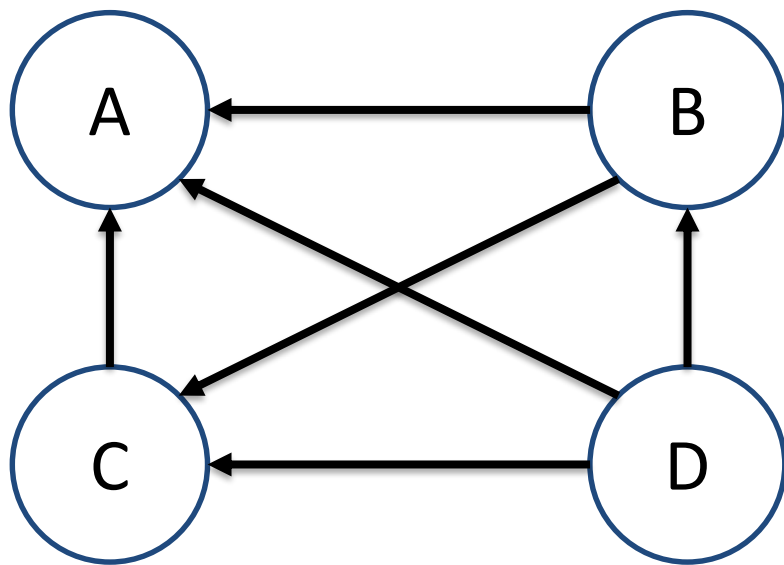
Чем больше ссылаются, тем лучше:

$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L(v)}$$



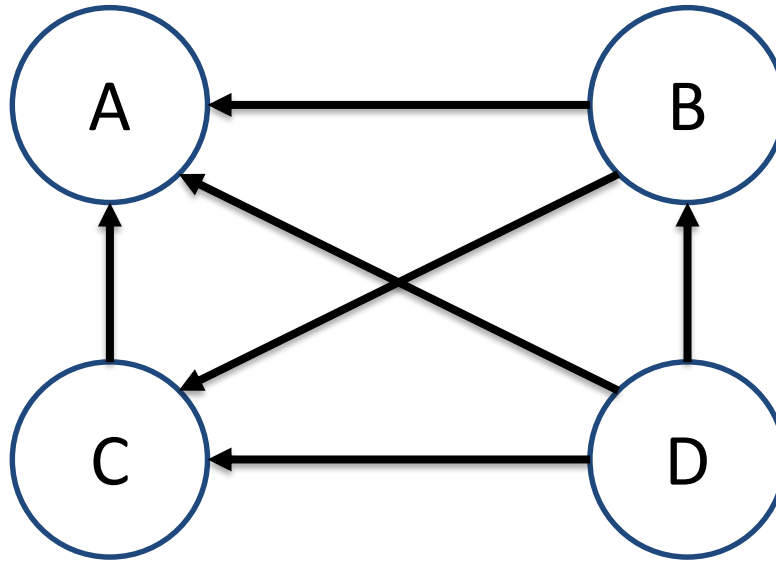
PageRank

1. Пользователь начинает обход «случайно»
2. Пользователь может «устать»



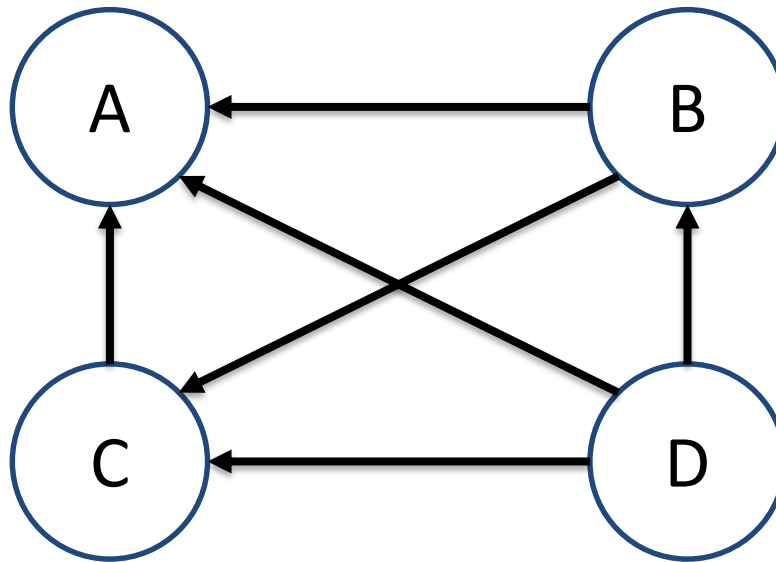
PageRank

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$



PageRank

1. Процесс сходится
2. Обычно требуется ~30-40 итераций



Google: PageRank + AdWords

PageRank не убил платную рекламу - он просто ее переместил

ПС нужен доход

Сайты продолжают платить, но уже за ранжирование в **рекламном блоке**

Баланс:

ПС получает прибыль за переходы, переходы происходят по релевантным объявлениям => удалять нерелевантные объявления и "плохой" контент

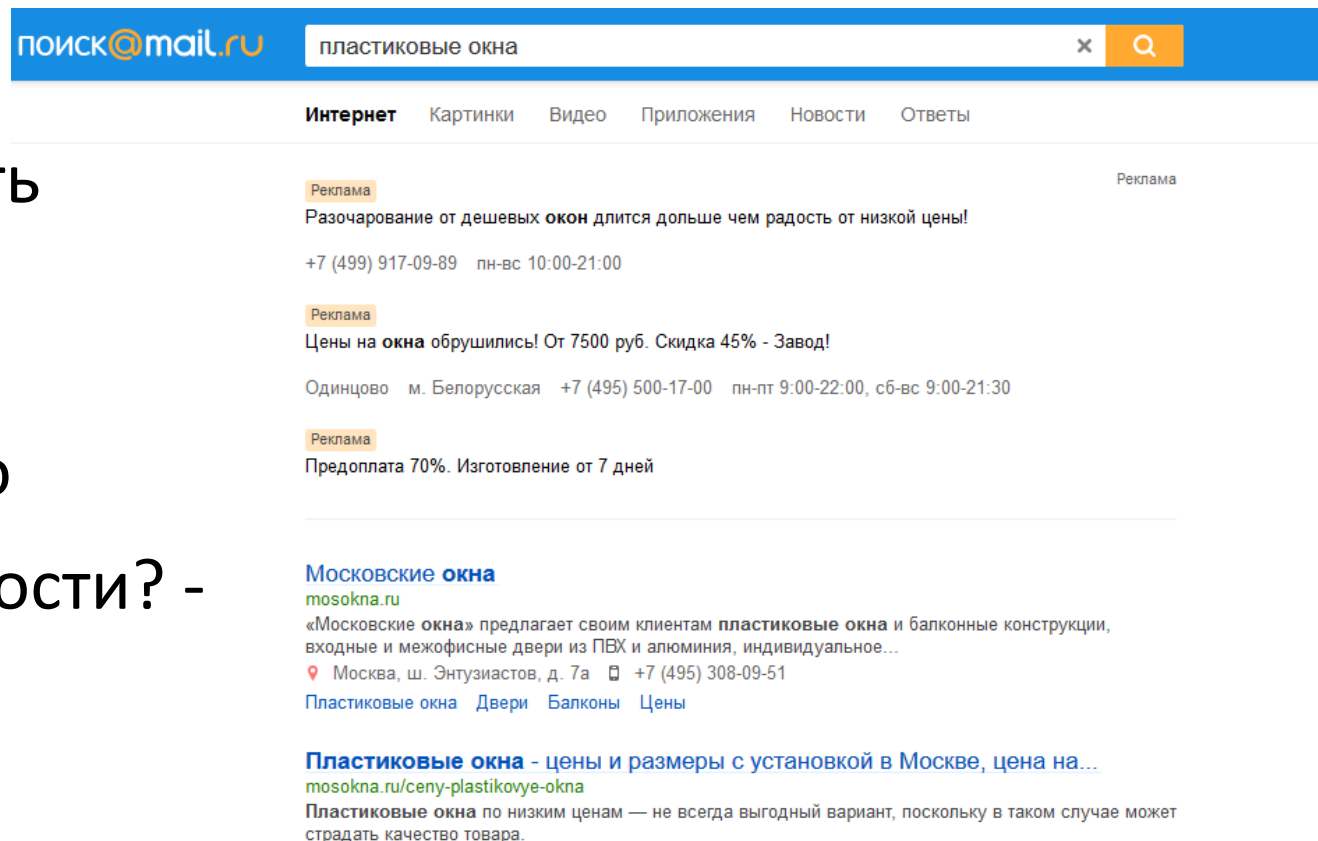
Реклама в ПС

Как ранжировать рекламу?

- по плате? - нерелевантно
- по релевантности? - невыгодно

Выход:

$$\text{rank} = f(\text{price}, \text{CTR})$$



поиск@mail.ru пластиковые окна x Q

Интернет Картинки Видео Приложения Новости Ответы

Реклама Реклама

Разочарование от дешевых окон длится дольше чем радость от низкой цены!
+7 (499) 917-09-89 пн-вс 10:00-21:00

Реклама

Цены на окна обрушились! От 7500 руб. Скидка 45% - Завод!
Одинцово м. Белорусская +7 (495) 500-17-00 пн-пт 9:00-22:00, сб-вс 9:00-21:30

Реклама

Предоплата 70%. Изготовление от 7 дней

Московские окна
mosokna.ru
«Московские окна» предлагает своим клиентам пластиковые окна и балконные конструкции, входные и межофисные двери из ПВХ и алюминия, индивидуальное...
📍 Москва, ш. Энтузиастов, д. 7а ☎ +7 (495) 308-09-51
[Пластиковые окна](#) [Двери](#) [Балконы](#) [Цены](#)

Пластиковые окна - цены и размеры с установкой в Москве, цена на...
mosokna.ru/ceny-plastikovye-okna
Пластиковые окна по низким ценам — не всегда выгодный вариант, поскольку в таком случае может страдать качество товара.

Second price auction

Аукцион Викри - однораундовый закрытый аукцион.
Победитель выплачивает вторую ставку

Закрытый => баланс между реальной оценкой и
максимально допустимыми затратами

Для рекламных систем - механизм Викри-Кларка-Гровса. N победителей, выплачивается сумма,
достаточная для удержания позиций

Second price auction

advertiser	bid	CTR			
A	4\$	0.01			
B	3\$	0.03			
C	2\$	0.06			
D	1\$	0.08			

bid - ставка за 1 переход с ПС

CTR - вероятность перехода:

$(\text{кол-во переходов}) / (\text{кол-во показов})$.

Это мера **релевантности**

Second price auction

advertiser	bid	CTR			
A	4\$	0.01			
B	3\$	0.03			
C	2\$	0.06			
D	1\$	0.08			

bid - ставка за 1 переход с ПС

CTR - вероятность перехода:

$(\text{кол-во переходов}) / (\text{кол-во показов})$.

Это мера **релевантности**

Как вычислить CTR, если объявление новое?

Second price auction

advertiser	bid	CTR	ad rank		
A	4\$	0.01	0.04		
B	3\$	0.03	0.09		
C	2\$	0.06	0.12		
D	1\$	0.08	0.08		

ad rank - фактическая полезность объявления для ПС.

Самая простая формула:

$$V(\text{bid}, \text{CTR}) = \text{bid} * \text{CTR}$$

Second price auction

advertiser	bid	CTR	ad rank	rank	
A	4\$	0.01	0.04 →	4	
B	3\$	0.03	0.09 →	2	
C	2\$	0.06	0.12 →	1	
D	1\$	0.08	0.08 →	3	

rank - итоговое ранжирование. Определяет, какую в итоге позицию занимает каждое из объявлений.

Сравним с GoTo

Second price auction

advertiser	bid	CTR	ad rank	rank	GoTo
A	4\$	0.01	0.04	4	1
B	3\$	0.03	0.09	2	2
C	2\$	0.06	0.12	1	3
D	1\$	0.08	0.08	3	4

rank - итоговое ранжирование. Определяет, какую в итоге позицию занимает каждое из объявлений.

Сравним с GoTo

Second price auction

advertiser	bid	CTR	ad rank	rank	price
A	4\$	0.01	0.04	4	
B	3\$	0.03	0.09	2	
C	2\$	0.06	0.12	1	
D	1\$	0.08	0.08	3	

price - сколько в итоге заплатит рекламодатель за каждый переход.

Second price auction

advertiser	bid	CTR	ad rank	rank	price
A	4\$	0.01	0.04	4	
B	3\$	0.03	0.09	2	
C	2\$	0.06	0.12	1	
D	1\$	0.08	0.08	3	

price - сколько в итоге заплатит рекламодатель за каждый переход.

$\text{adRank}_1 > \text{adRank}_2$

$V(\text{price}_1, \text{CTR}_1) > V(\text{bid}_2, \text{CTR}_2)$

Second price auction

advertiser	bid	CTR	ad rank	rank	price
A	4\$	0.01	0.04	4	
B	3\$	0.03	0.09	2	
C	2\$	0.06	0.12	1	
D	1\$	0.08	0.08	3	

price - сколько в итоге заплатит рекламодатель за каждый переход.

$$V(\text{price}_1, \text{CTR}_1) > V(\text{bid}_2, \text{CTR}_2)$$

$$\text{price}_1 * \text{CTR}_1 > \text{bid}_2 * \text{CTR}_2$$

$$\text{price}_1 * \text{CTR}_1 = \text{bid}_2 * \text{CTR}_2 (+0.01\$)$$

$$\text{price}_1 = \text{bid}_2 * \text{CTR}_2 / \text{CTR}_1 (+0.01\$)$$

Second price auction

advertiser	bid	CTR	ad rank	rank	price
A	4\$	0.01	0.04	4	
B	3\$	0.03	0.09	2	
C	2\$	0.06	0.12	1	
D	1\$	0.08	0.08	3	

$\text{price}_1 = \text{bid}_2 * \text{CTR}_2 / \text{CTR}_1 (+0.01\$)$

1: C, bid=2\$, CTR=0.06

2: B, bid=3\$, CTR=0.03

Second price auction

advertiser	bid	CTR	ad rank	rank	price
A	4\$	0.01	0.04	4	
B	3\$	0.03	0.09	2	2.68\$
C	2\$	0.06	0.12	1	1.51\$
D	1\$	0.08	0.08	3	0.51\$

$$\text{price}_1 = \text{bid}_2 * \text{CTR}_2 / \text{CTR}_1 (+0.01\$)$$

Second price auction

advertiser	bid	CTR	ad rank	rank	price
A	4\$	0.01	0.04	4	0.01\$
B	3\$	0.03	0.09	2	2.68\$
C	2\$	0.06	0.12	1	1.51\$
D	1\$	0.08	0.08	3	0.51\$

$$\text{price}_1 = \text{bid}_2 * \text{CTR}_2 / \text{CTR}_1 (+0.01\$)$$

Качество поиска

Качество поиска

- Релевантность
- Покрытие многозначности
- Простота UI
- Уменьшение ошибок пользователя
- Полнота

Качество поиска. Полнота

Большинству пользователей не нужна.

Проявляется на непопулярных запросах.

Найдите документы со словом
собакокрадство

Качество поиска

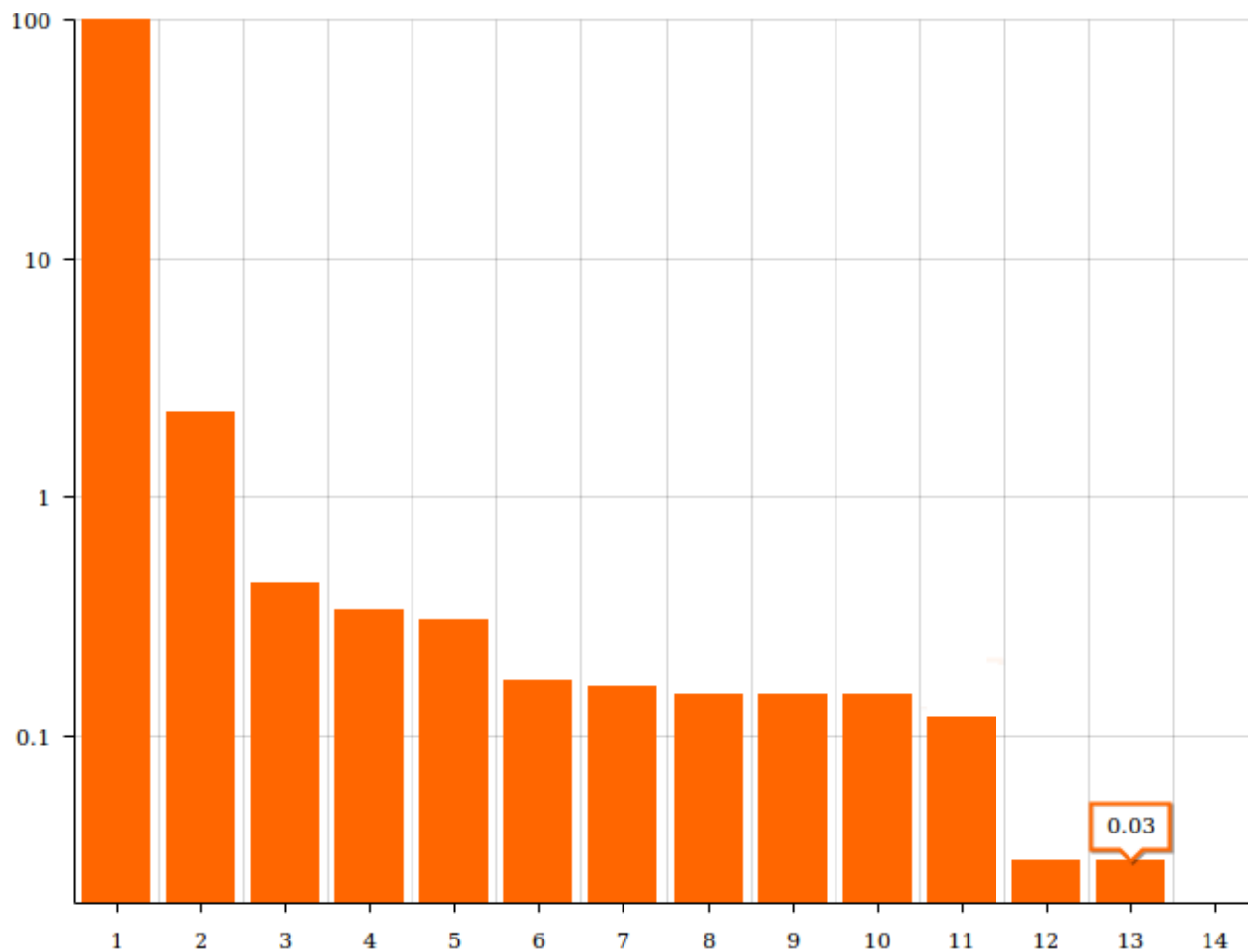
Полнота по непопулярным запросам - способ оценить размер полезного индекса.

<http://www.analyzethis.ru/?analyzer=rare&date=2016-09-01&lang=ru&location=ru>

Независимые метрики различных аспектов ПС:

<http://www.analyzethis.ru/>

Пользователи и поиск. Как далеко заходят пользователи.



Поисковый спайдер



Задача:

Нужно скачать 1 сайт.

Ваши предложения?

Задача:

Нужно скачать 1 сайт.

- Нужно учитывать допустимую нагрузку:
 - Качать в 1 поток
 - Выдерживать паузы между запросами

Проблема:

Сайтов много

Страниц еще больше

Времени мало

Спайдер

1. Постановка задачи
2. Выкачка
3. Обновление
4. Хранение

Требования к спайдеру

1. Politeness
2. Freshness
3. Actuality
4. Производительность
5. Масштабируемость

URL

RFC: <https://www.ietf.org/rfc/rfc1738.txt>

<http://site.ru/path?page=10>

http - схема

site.ru - хост

path - путь

page=10 - query

IP

Уникальный адрес сетевого узла

```
$ host go.mail.ru
```

```
$ host ru.wikipedia.org
```

DNS

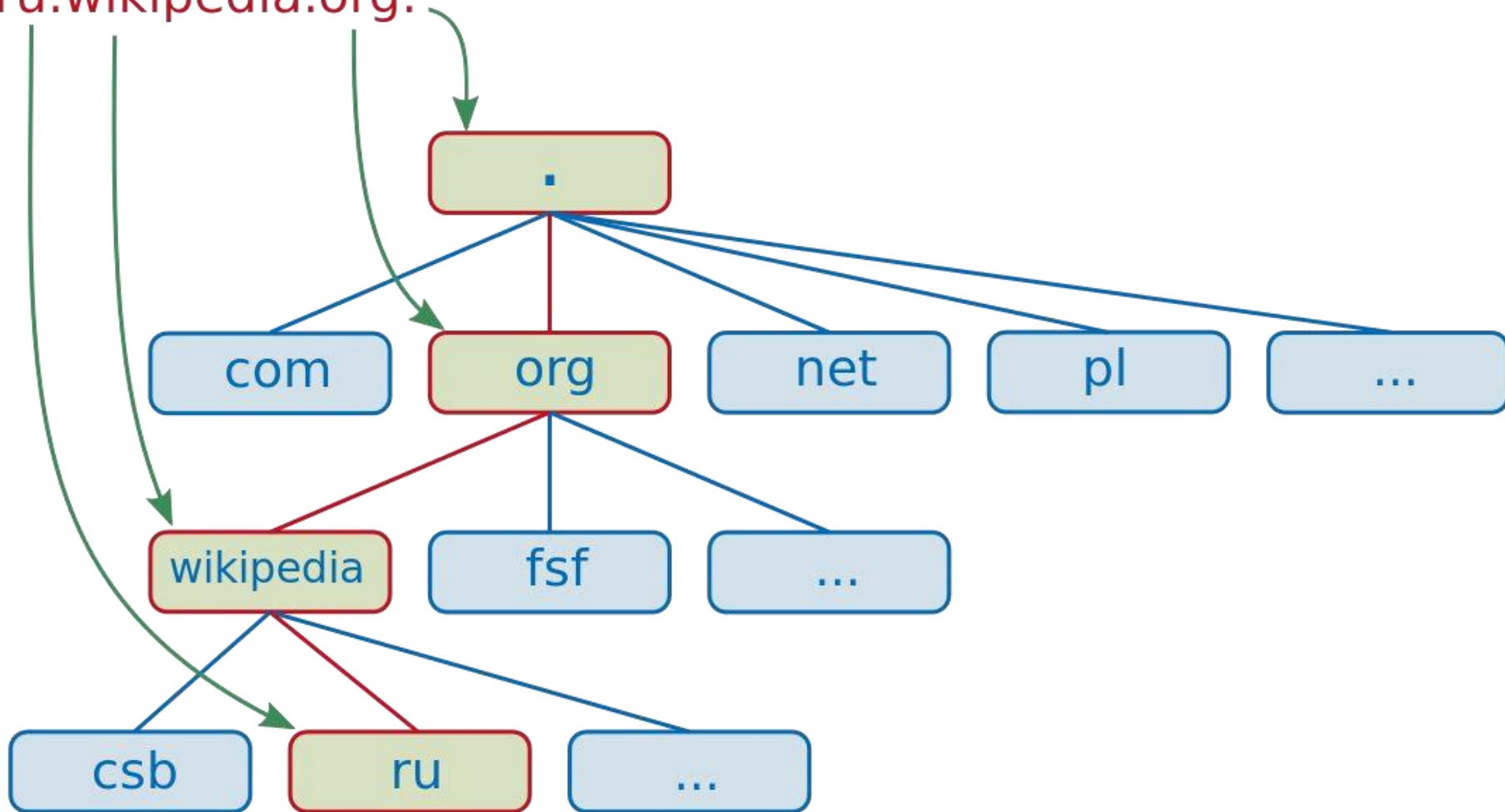
DNS - сервис для получения информации о доменах. Нам нужна информация об IP.

url -> ip

информация предоставляется иерархической системой серверов - может быть долго

DNS

ru.wikipedia.org.



DNS

```
$ dig -t NS .
```

```
$ dig -t NS @e.root-servers.net. ru
```

```
...
```

```
$ dig -t A @???? go.mail.ru
```

Сколько ір-адресов у сайта?

Сколько ip-адресов у сайта?

1. 1-1:

```
$ host -v -t A zonova.xyz
```

2. 1-n: снижение нагрузки (для высоконагруженных систем)

```
$ host -v -t A go.mail.ru
```

3. m-1: снижение стоимости

```
$ host -v -t A catalogr.ru
```

```
$ host -v -t A redbook73.ru
```

Robots.txt

```
User-agent: *  
Crawl-delay: 50  
Disallow: /admin  
Allow: /article
```

Хорошие роботсы:

<http://lenta.ru/robots.txt>

Плохие роботсы:

<https://money.yandex.ru/robots.txt>

Robots.txt

```
User-agent: *  
Crawl-delay: 50  
Disallow: /admin  
Allow: /article
```

Какие из этих документов
можно качать?

<http://site.ru/>

<http://site.ru/admin>

<http://site.ru/admin/article>

<http://site.ru/article/admin>

<http://site.ru/post>

Robots.txt

```
User-agent: *  
Crawl-delay: 50  
Disallow: /admin  
Allow: /article
```

Какие из этих документов
можно качать?

<http://site.ru/>

<http://site.ru/admin>

<http://site.ru/admin/article>

<http://site.ru/article/admin>

<http://site.ru/post>

Спайдер

1. Постановка задачи
2. Выкачка
3. Обновление
4. Хранение

Алгоритм

1. "Точка входа" - seed-урлы
2. Скачали
3. Распарсили, извлекли урлы, отправили урлы в очередь на обкачку
4. goto #2

Seed-урлы

КАТАЛОГ@mail.ru®

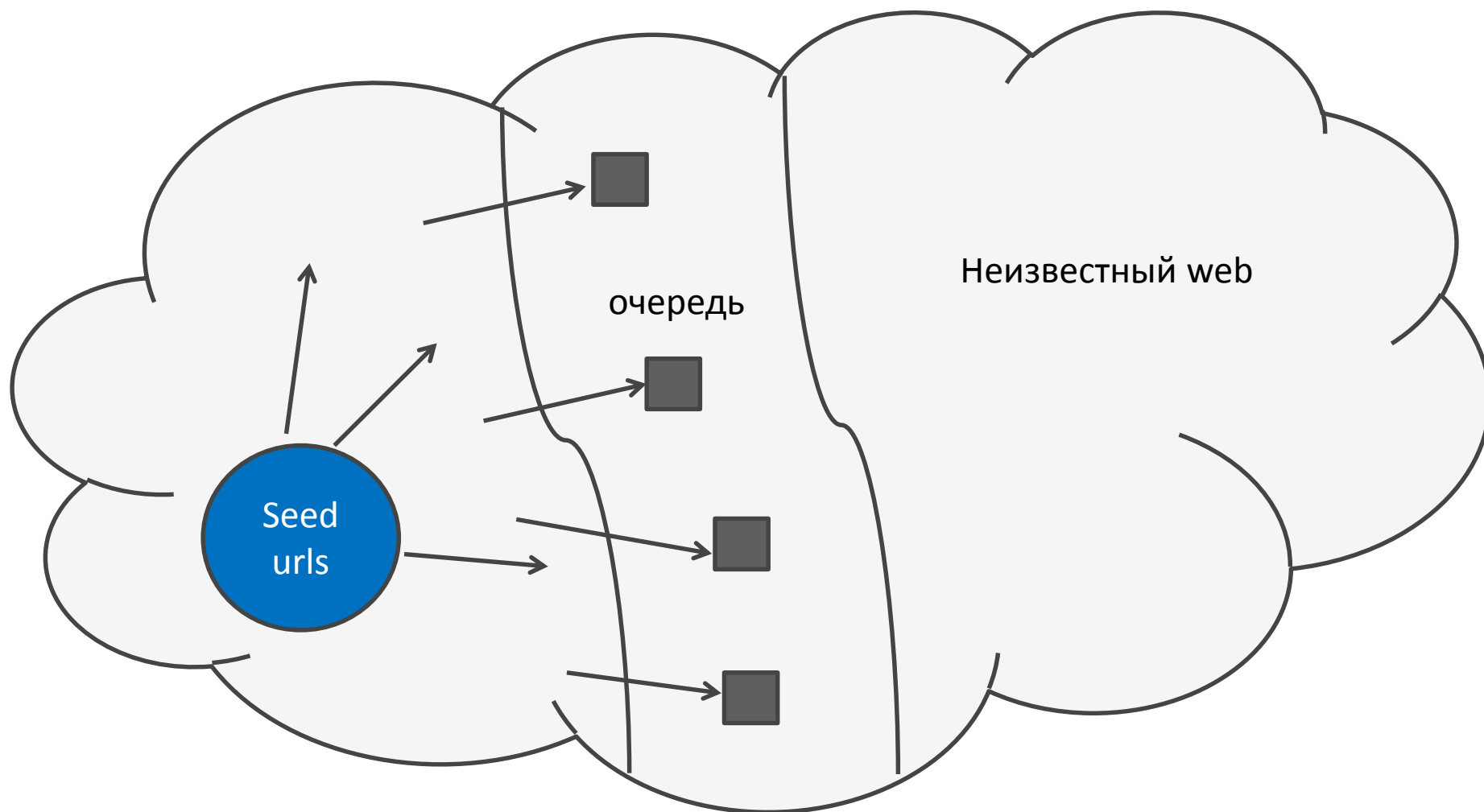
Яндекс
каталог

РЕЙТИНГ@mail.ru



Википедия
Свободная энциклопедия

Выкачка



Ответы сервера

Какие бывают?

2xx - успешно

3xx - перенаправление

4xx - ошибка клиента

5xx - ошибка сервера

Особенности контента

1. Тип контента
2. Кодировка

Тип контента

html, jpeg, pdf, xml, mp3 и т.д.

Как определить:

1. Content-Type: text/html
2. По первым символам контента

```
1 <!DOCTYPE html>  
2 <html>  
3 <head>
```

Не всё так просто:

<http://kiev-ehudi.org.ua/>

БНОПНЯ

\$ echo БНОПНЯ | iconv -f CP1251 -t KOI8R



Какая кодировка?

Не надо быть умнее браузера.

1. Content-type: charset в http-head

```
$ wget --spider -Sq https://en.wikipedia.org/wiki/Sicily  
2>&1 | grep charset
```

```
Content-Type: text/html; charset=UTF-8
```

Какая кодировка?

Не надо быть умнее браузера.

1. Content-type: charset в http-head
2. Meta-charset

```
$ wget -SO ch1 http://solarboat.ru/catalog/lodki_solar/ 2>&1 | grep charset
```

```
Content-Type: text/html; charset=windows-1251
```

```
$ grep charset ./ch1
```

```
<meta http-equiv="Content-Type" content="text/html;  
charset=windows-1251" />
```


Какая кодировка?

Не надо быть умнее браузера.

1. Content-type: charset в http-head
2. Meta-charset

Определите кодировку:

<http://www.emalirovka-vann.ru/>

<http://ievpdgh.22web.org/?i=1>

Какая кодировка?

<http://www.emalirovka-vann.ru/>

Http-head: cp1251

Meta: utf8

<http://ievpdgh.22web.org/?i=1>

Http-head: -

Meta: utf8;cp1251

Какая кодировка?

<http://www.emalirovka-vann.ru/>

Http-head: cp1251

Meta: utf8

Res: utf8

<http://ievpdgh.22web.org/?i=1>

Http-head: -

Meta: utf8;cp1251

Res: utf8

Извлечение ссылок (discovering)

```
<a href="...">
```

Помним о politeness:

```
<meta name="robots" content="nofollow" />
```

```
<a href="signin.php" rel="nofollow">Войти</a>
```

Извлечение ссылок (discovering)

Ссылки бывают:

1. Внутренние и внешние
2. Абсолютные и относительные
3. Валидные и невалидные

Минутка прекрасного:

<http://www.mongolianembassy.ru/>

Абсолютные и относительные ссылки

<http://site.ru/page/1>

`` --> <http://site.ru/page/2>

`` --> <http://site.ru/2>

`` --> <http://site.ru/d3>

`` --> <http://site.com/page>

`` --> <http://abc.org/g>

Нельзя брать все ссылки



Нельзя брать все ссылки

1. Robots.txt
2. Некоторые документы мы уже качали
3. Внутренний blacklist:
 1. Правильные ограничения: <http://go.mail.ru/robots.txt>
 2. <https://www.iconfinder.com/search/?q=search>

А еще сайты могут быть "бесконечными":

<http://www.calend.ru/day/1-2-2050/>

Что брать и сколько?

Решает внешняя задача - scheduler

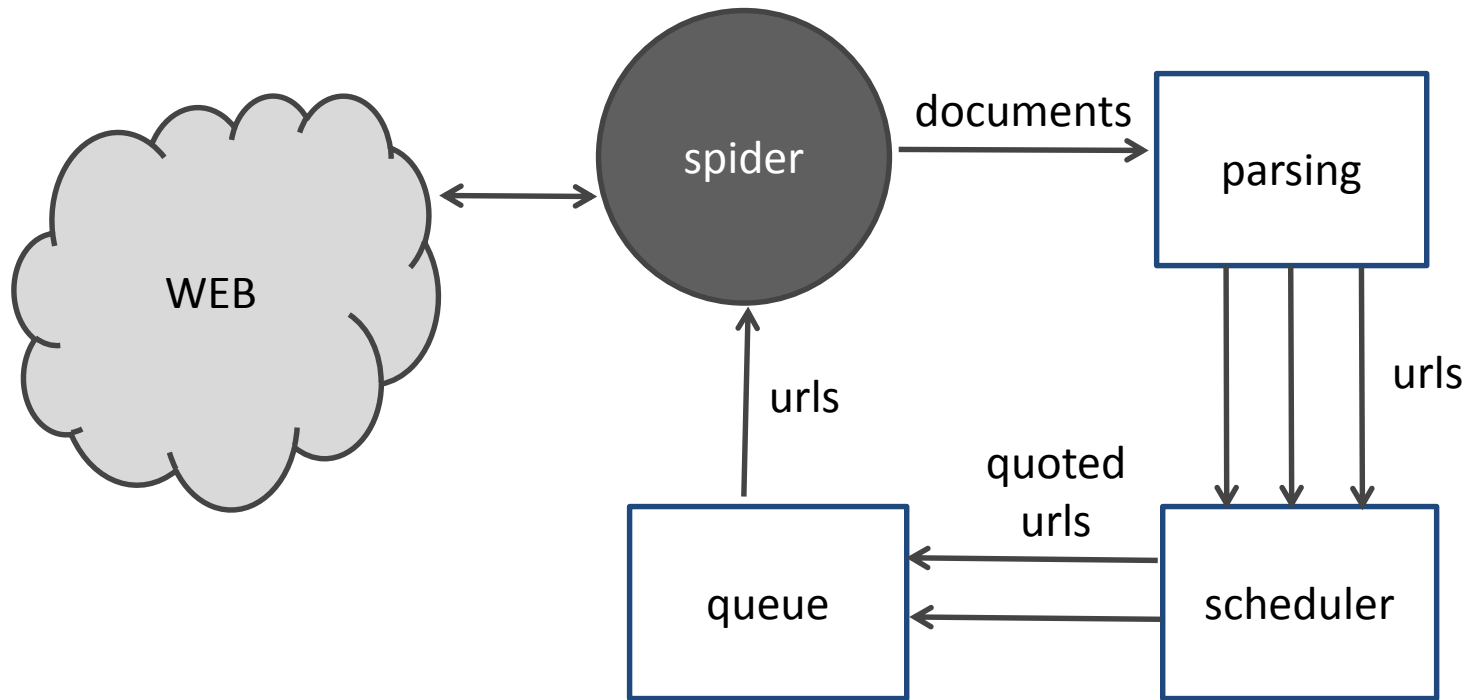
Учитывает:

1. Количество уже скачанных документов с сайта (успешно и нет)
2. Свойства скачанных документов (тип / язык)
3. Свойства самого сайта (посещаемость, CTR и т.д.)

Формируется квота.

Лекции про ранжирование

Spider & utils



Спайдер

1. Постановка задачи
2. Выкачка
3. Обновление
4. Хранение

Зачем перекачивать страницы?

1. Обновилось содержимое
2. Появились ссылки на новые страницы

Пример: главная страница сайта

Как часто перекачивать?

Простой подход:

если страница изменилась - $T = T/2$

если страница не изменилась - $T = T*2$

Усложнение:

- История выкачки
- Ранк сайта

Что важнее?

Выкачка новых страниц или перекачка старых?



Как понять, что страница изменилась?

Как понять, что страница изменилась?

<http://lenta.ru/>

<http://wellclix.net/>

<https://www.adme.ru/>

Как понять, что страница изменилась?

1. Брать только "чистый" контент
2. Удаление обвязки

Об этом - в другой лекции

Как понять, что страница изменилась?

Вэбмастера в одной лодке с нами

Http-response:

eTag

Last-Modified

В основном - для статического контента

Как понять, что страница изменилась?

```
$ HEAD http://s.imgur.com/images/loaders/ddddd1_181817/24.gif
200 OK
ETag: "57a25124-14f9"
Last-Modified: Wed, 03 Aug 2016 20:16:36 GMT
```

Как понять, что страница изменилась?

```
$ HEAD http://s.imgur.com/images/loaders/ddddd1_181817/24.gif  
200 OK
```

```
ETag: "57a25124-14f9"
```

```
Last-Modified: Wed, 03 Aug 2016 20:16:36 GMT
```

```
$ HEAD -H 'If-None-Match: "57a25124-14f9"'  
http://s.imgur.com/images/loaders/ddddd1_181817/24.gif  
304 Not Modified
```

```
$ HEAD -H 'If-None-Match: "57a25124-14f8"'  
http://s.imgur.com/images/loaders/ddddd1_181817/24.gif  
200 OK
```

```
$ HEAD -H 'If-Modified-Since: Wed, 03 Aug 2016 20:16:36 GMT'  
http://s.imgur.com/images/loaders/ddddd1_181817/24.gif  
304 Not Modified
```

Дополнительные источники информации

AliWeb - поисковик, который использовал заранее подготовленные "индексные файлы", содержащие список урлов и их описание (по усмотрению владельца ресурса)

А сейчас?

Дополнительные источники информации

<http://simonscat.tumblr.com/rss>

```
<?xml version="1.0" encoding="UTF-8"?>
<rss xmlns:dc="http://purl.org/dc/elements/1.1/" version="2.0">
<channel>
  <description>Channel description</description>
  <title>Simon's Cat</title>
  <item>
    <title>Simon's Cat refusing to face Monday! </title>
    <description>post description</description>
    <link>http://simonscat.tumblr.com/post/150306700829</link>
    <pubDate>Mon, 12 Sep 2016 12:33:35 +0100</pubDate>
  </item>
  ...
</channel>
```

Дополнительные источники информации

<http://all-t-shirts.ru/sitemap.xml?start=0>

```
<urlset>
```

```
  <url>
```

```
    <loc>http://all-t-shirts.ru/</loc>
```

```
    <lastmod>2016-03-28T00:03:15+03:00</lastmod>
```

```
    <changefreq>daily</changefreq>
```

```
  </url>
```

```
  ...
```

```
</urlset>
```

Спайдер

1. Постановка задачи
2. Выкачка
3. Обновление
4. Хранение

Хранение скачанных документов

Ваши варианты?

Хранение скачанных документов

Документ <--> урл

Ключ - f(url)

Практика. Есть разные способы записать один URL

<https://ru.wikipedia.org/wiki/%D0%9F%D0%BE%D0%BD%D0%B8>

<https://ru.wikipedia.org/wiki/Пони>

<https://ru.wikipedia.org/wiki/%CF%EE%ED%E8>

http://kikolani.com/blog-post-promotion-ultimate-guide?utm_source=kikolani&utm_medium=320banner&utm_campaign=bpp

<http://kikolani.com/blog-post-promotion-ultimate-guide>

<http://scifi.stackexchange.com/questions?page=4&sort=newest>

<http://scifi.stackexchange.com/questions?sort=newest&page=4>

<https://music.yandex.ru/album/3575649/track/29692077>

<http://music.yandex.ru/album/3575649/track/29692077/>

<https://www.music.yandex.ru/album/3575649/track/29692077>

http://opennet.ru/docs/RUS/inet_book/4/45/retr4514.html

http://www.opennet.ru/docs/RUS/inet_book/4/45/retr4514.html

<http://домены.рф/>

<http://xn--d1acufc5f.xn--p1ai/>

<http://domeny.rf/>

Хранение документов

Нормализация урла

RFC: <https://www.ietf.org/rfc/rfc1738.txt>

Хранение документов

И проверка на валидность

<http://domeny.rf/> - .rf не существует

Хранение документов

Нормализованный URL - всегда в ASCII

Percent-encoding для query и пути

```
$ python -c "import urllib, sys; print urllib.quote(sys.argv[1])" Пони  
%D0%9F%D0%BE%D0%BD%D0%B8
```

Punycode для имени домена:

```
$ python -c "import urllib, sys; print sys.argv[1].decode('utf-8').encode('idna')"  
домены.рф  
xn--d1acufc5f.xn--p1ai  
$ python -c "import urllib, sys; print sys.argv[1].decode('idna')" xn--d1acufc5f.xn--  
p1ai  
домены.рф
```

Хранение документов

Нормализованный URL - всегда в ASCII

<https://ru.wikipedia.org/wiki/%D0%9F%D0%BE%D0%BD%D0%B8>

<https://ru.wikipedia.org/wiki/Пони>

<https://ru.wikipedia.org/wiki/%CF%EE%ED%E8>

<http://домены.рф/>

<http://xn--d1acufc5f.xn--p1ai/>

Хранение документов

utm-метки для маркировки траффика

Параметры, которые игнорируются сервером, но учитываются в статистике

Позволяют оценить успешность рекламных кампаний (источники переходов)

Хранение документов

utm-метки для маркировки траффика

http://kikolani.com/blog-post-promotion-ultimate-guide?utm_source=kikolani&utm_medium=320banner&utm_campaign=bpp

<http://kikolani.com/blog-post-promotion-ultimate-guide>

Хранение документов

www. - наследие старого мира

Большинство - редиректят на нужную версию

Есть исключения:

www.music.yandex.ru - редиректит на корневик

[http://www.opennet.ru/](http://www.opennet.ru) и [http://opennet.ru/](http://opennet.ru) - обе отдают контент (одинаковый)

Хранение документов

www. - наследие старого мира

<https://music.yandex.ru/album/3575649/track/29692077>

<http://music.yandex.ru/album/3575649/track/29692077/>

<https://www.music.yandex.ru/album/3575649/track/29692077>

http://opennet.ru/docs/RUS/inet_book/4/45/retr4514.html

http://www.opennet.ru/docs/RUS/inet_book/4/45/retr4514.html

Хранение документов

Зеркало - сайт (до 80%) дублирующий контент оригинала

1. Защита от падения
2. ... и от блокировок (lurkmore.to, lurklurk.com, lurkmirror.ml)
3. Дорогой внешний трафик - локальное зеркало

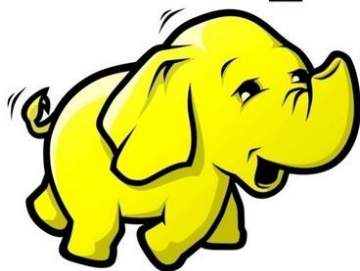
Как бороться? Искать дубликаты (другая лекция)

Хранение документов

> 40 Pb

> 100 млрд. документов

hadoop



Хранение документов



←EROSPIKE→



Вопросы?