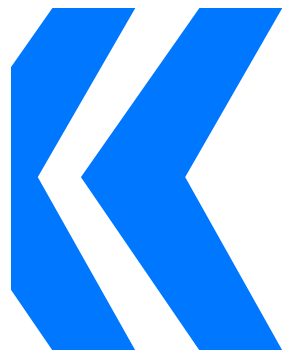


Поисковый робот

Поиск дубликатов

Информационный поиск. Лекция No2





Обо мне

Сергей Чуриков

Программист в группе инфраструктуры
поиска VK

Содержание

Поисковой робот

Постановка задачи

Выкачка

Обновление

Хранение

Поиск дубликатов

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

Необходимо качать документы

- Чтобы было среди чего производить поиск, документы должны попасть к нам в базу
- Для этого нужен спайдер - система которая будет обходить интернет в поисках релевантных страниц



УПРОЩЕННЫЙ АЛГОРИТМ РАБОТЫ ПОИСКОВОЙ СИСТЕМЫ

Постановка задачи

Задача

Нужно скачать сайт.

Какие могут быть проблемы и как качать?

Требования к спайдеру

- Вежливость.
- Качество и свежесть обкачиваемых страниц.
- Производительность.
- Масштабируемость.
- Устойчивость к “ловушкам” - разные версии для робота и пользователя, скрытые ссылки для робота, циклы.

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.

В том числе сопоставление URL -> IP.

Скільки 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

Robots.txt

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

Примеры роботов:
<http://lenta.ru/robots.txt>
<https://yandex.ru/robots.txt>
<https://tnt-online.ru/robots.txt>

RFC - <https://datatracker.ietf.org/doc/html/rfc9309>.

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>

Выкачка

Выкачка

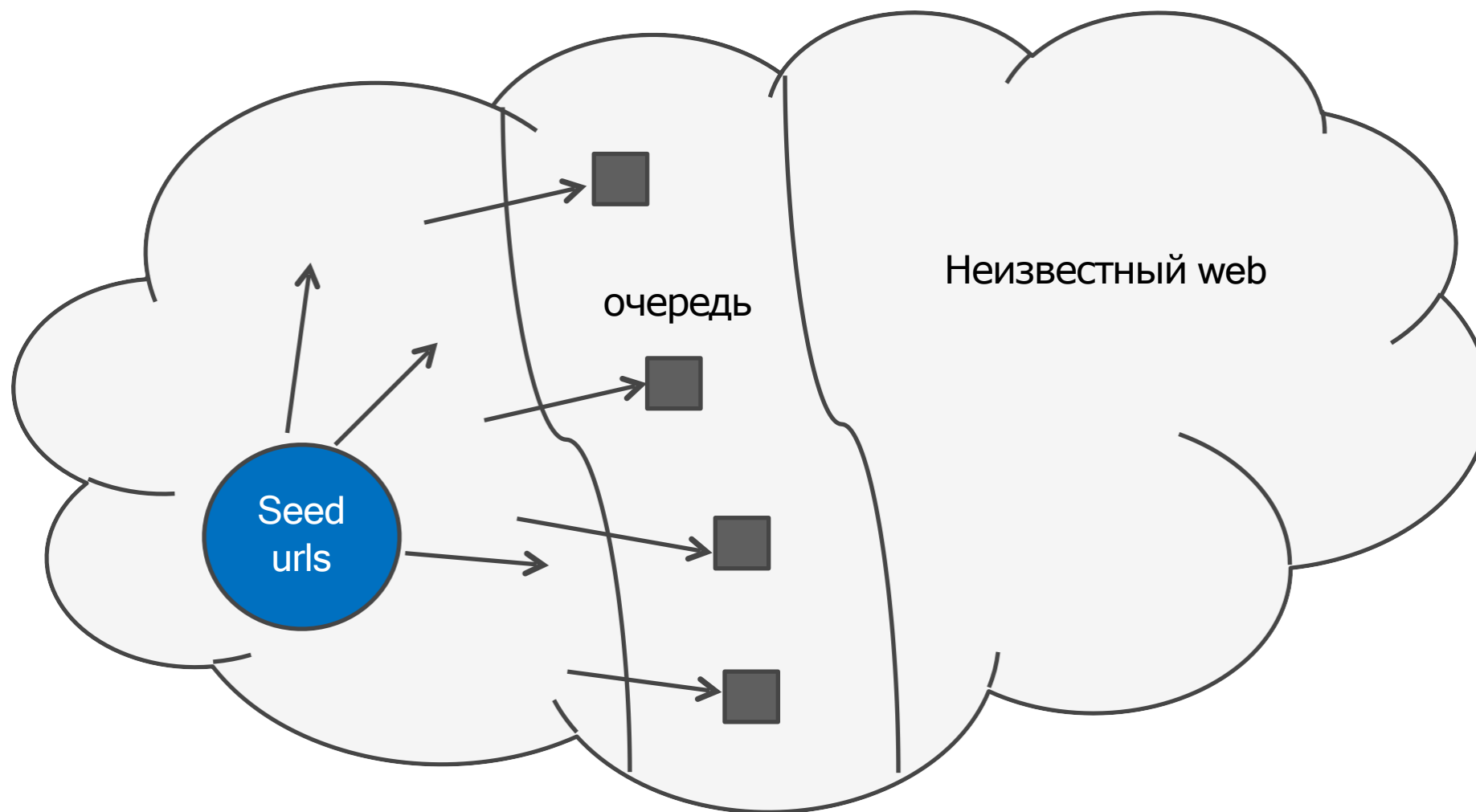
Алгоритм:

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

Определяем seed-урлы

- Каталоги сайтов.
- Википедия.
- ...

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. Заголовок Content-Type. В первую очередь интересен text/html.
2. По первым символам контента.

Какая кодировка?

Не надо быть умнее браузера. «Чем ближе к тексту, тем правильнее»

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 index.html https://yuhui-lin.github.io/post/  
2021-06-01_clickhouse-json/ 2>&1 | grep charset
```

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

```
$ grep charset ./index.html
```

```
<meta charset="utf-8" />
```

Какая кодировка?

Не надо быть умнее браузера.

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

Какая кодировка?

Http-head: utf8

Meta: cp1251

Http-head: -

Meta: utf8;cp1251

Какая кодировка?

Http-head: utf8

Meta: cp1251

Res: utf8

Http-head: -

Meta: utf8;cp1251

Res: utf8

Извлечение ссылок (discovering)

``

Помним о politeness:

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

`Войти`

Извлечение ссылок (discovering)

Ссылки бывают:

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

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

1. Robots.txt
2. Некоторые документы мы уже качали
3. Внутренний blacklist:
 1. Правильные ограничения: <https://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?

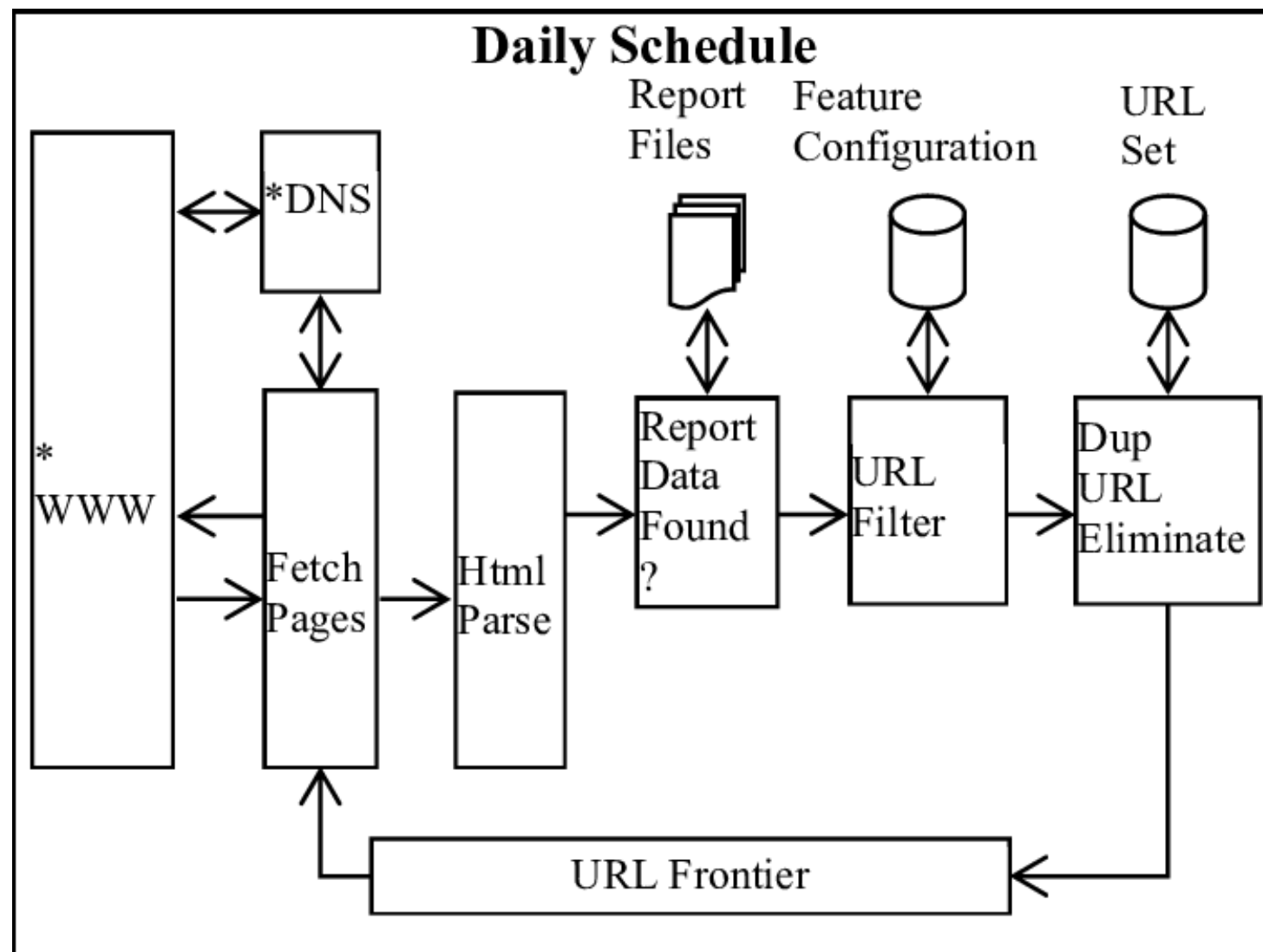
Решает внешняя задача - scheduler

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.

Spider - Архитектура



Обновление

Обновление

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 - $T = T*2$

Еще идеи?

Как часто перекачивать?

Простой подход:

если страница изменилась - $T = T/2$

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

Усложнение:

- История выкачки
- Ранк сайта
- Интервалы обкачки

Что важнее?

Выкачка новых страниц
или перекачка старых?



Как понять, что страница изменилась?

Как понять, что страница изменилась?

<https://dzen.ru/news>

<https://adme.media/>

<https://lenta.ru/>

Как понять, что страница изменилась?

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

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: "f49abbb822e99d5e1d1e7020daeea5be"

Last-Modified: Thu, 15 Feb 2024 17:47:38 GMT

Как понять, что страница изменилась?

```
$ HEAD http://s.imgur.com/images/loaders/ddddd1\_181817/24.gif
```

```
200 OK
```

```
ETag: "f49abbb822e99d5e1d1e7020daeea5be"
```

```
Last-Modified: Thu, 15 Feb 2024 17:47:38 GMT
```

```
$ HEAD -H 'If-None-Match: "f49abbb822e99d5e1d1e7020daeea5be"' http://s.imgur.com/images/loaders/ddddd1\_181817/24.gif
```

```
304 Not Modified
```

```
$ HEAD -H 'If-None-Match: "asd"' http://s.imgur.com/images/loaders/ddddd1\_181817/24.gif
```

```
200 OK
```

```
$ HEAD -H 'If-Modified-Since: Tue, 20 Feb 2024 15:59:09 GMT' http://s.imgur.com/images/loaders/ddddd1\_181817/24.gif
```

```
304 Not Modified
```

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, 02 Feb 2024 13:53:00 +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>2023-03-06 20:26:18</lastmod>
```

```
    <changefreq>yearly</changefreq>
```

```
    <priority>0.5</priority>
```

```
  </url>
```

```
  ...
```

```
</urlset>
```

Хранение

Хранение скачанных документов

Как и где будем хранить ?

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

```
$ python3 -c "import urllib.parse, sys; print(urllib.parse.quote(sys.argv[1]))" Пони  
%D0%9F%D0%BE%D0%BD%D0%B8
```

Punycode для имени домена:

```
$ python3 -c "import sys; print(sys.argv[1].encode('idna'))" домены.рф  
b'xn--d1acufc5f.xn--p1ai'
```

```
$ python3 -c "import sys; print(sys.argv[1].encode().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?utm_source=kikolani&utm_medium=320banner&utm_campaign=bpp)

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

Хранение документов

Сортировка get параметров

<http://kikolani.com/blog-post-promotion-ultimate-guide?b=1&c=2&a=3>

<http://kikolani.com/blog-post-promotion-ultimate-guide?a=3&b=1&c=2>

Get параметры могут быть важны для ответа сайта

Хранение документов

[www.](#) - наследие старого мира

Большинство - редиректят на нужную версию

Но бывают исключения

Хранение документов

<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

Хранение документов

- > Десятки Pb
- > Сотни млрд. документов

Технологии

- Hadoop (MapReduce & Spark)
- Hbase

Хранение документов

- > Десятки Pb
- > Сотни млрд. документов

Технологии

- Hadoop (MapReduce & Spark)
- Hbase

Поиск дубликатов

Дубликаты



Дубликаты

[illegible][illegible]

[>>> www.zigzag.ru/animals/kapibara/](#) | Слово капибара (Hydrochoerus hydrochaeris).html
 < Назад | Домашние животные / Капибары, или водосвинки (Hydrochoerus hydrochaeris) |

Алфавитный указатель

А Б В Г Д Е Ё Ж З И Й К Л М Н О П Р С Т У Ф Х Ц Ч Ш Щ Э Ю Я

Капибара, или водосвинка (*Hydrochoerus hydrochaeris*) – полуводное травоядное млекопитающее из семейства водосвиновых (Hydrochoridae), единственное представительство в семействе. Капибара – самый крупный род современных грызунов. На языке индейцев гуарани слово капибара означает Господин Траур.

Внешний вид

Длина тела взрослой капибары достигает 1-1,35 м, высота в холке - 50-60 см. Самцы весят 34-63 кг, а самки - 36-65,5 кг. Самки, как правило, крупнее самцов.

Телосложение тяжёлое. Внешне капибара напоминает гигантскую бобровую морскую свинку. Голова крупная, массивная с широкой, тупой мордой. Верхняя губа толстая. Уши короткие, округлые. Ноздри широко расставлены. Глаза маленькие, расположены высоко на голове и отнесены несколько назад. Хвост rudimentary. Колючести довольно короткие; передние – 4 пальца (пальцы было бы шарики), задние – 3 палачи. Пальцы соединены небольшими плавающими перепонками и снабжены короткими сильными когтями. Тело покрыто длинными (30-120мм) и жесткими волосами; подшерсток отсутствует. Окрас верхней стороны тела от рыжеватого-бурого до сероватого, бурый, как правило, желтовато-бурый. Молодичек окрашены светлее. У половозрелых самцов на верхней части морды расположен участок кожи с многочисленными глубокими салыми железами. У самок имеется 6 пар борошковых сосков.


Череп массивный, с широкими и сильными скуловыми дугами. Зубов 20. Щечные зубы без корней, растут в течение всей жизни животного. Резцы широкие, имеют продольную бороздку на наружной поверхности. Малая и большая берцовые кости частично срastаются между собой. Ключицы нет. Хромосом в диплоидном наборе 66.

Вот как описывает капибару Джеральд Дэррил в "Трёх билетах до Эдиенге": "Этот гигантский грызун представляет собой жирного зверька с продолговатым телом, покрытым жёсткой помятой шерстью густой коричневой шерсти. Передние лапы у капибары длиннее задних, массивный орудок не имеет хвоста, и поэтому у неё всегда такой вид, будто она вот-вот соберётся сесть. У неё крупные лапы с широкими перепончатыми пальцами, а ногти на передних лапах, короткие и тупые, удивительно напоминают миниатюрные копытца. Вид у неё весьма аристократический: её плоская широкая голова и тупая, почти квадратная морда имеют благообразно-покровительственное выражение, предположительно сходство с лодочным лысом. По земле капибара передвигается характерной шарканью походкой или скандет взрывчатому гапотам, в воде же плывлет и ныряет с поразительной лёгкостью и проворством. Капибара - флегматичный добродушный вегетарианец, лишённый ярких индивидуальных черт, присущих некоторым его соратникам, но этот недостаток восполняется у неё спокойным и дружелюбным нравом."

Распространение и среда обитания

Капибара встречается по берегам разнообразных водоёмов в тропических и умеренных частях Центральной и Южной Америки, восточне Анд - от Панамы до Уругвая и северо-востока Аргентины (до 38°17' ю. ш., провинция Бунос-Айрес).

Дубликаты



cyclowiki.org

Навигация

- Помощь Циклопедии
- Сообщить об ошибке
- Форум
- FAQ
- Формат статей
- Качественные статьи
- Случайная статья
- Новые страницы
- Свежие правки

Инструменты

- Ссылка сюда
- Связанные правки
- следа правки
- Версия для печати
- Постоянная ссылка

Статья | Обсуждение

Капибара

[Читая](#) | [Правка](#) | [История](#) | [Поиск](#)


Капиба́ра, или **водосви́нка** (лат. *Hydrochoerus hydrochaeris*) — полуводное травоядное млекопитающее из семейства водосвинковых (Hydrochoeridae), единственный представитель в семействе.

Капибара — самый крупный среди современных грызунов.

Содержание [убрать]

- Внешний вид
- Происхождение и разновидности
- Ареал
- Образ жизни
 - Окружающая среда
 - Сообщество
 - Размножение
 - Питание
 - Болезни
 - Содержание в неволе
 - Продолжительность жизни
- Охрана и статус вида
- Популярность
- Интересные факты
- Источники
- Литература
- Ссылки

Внешний вид



Капибара — это водосвинка, самый крупный современный грызун в мире. Длина тела капибары достигает полутора метров, вес — шестидесяти килограмм. Животное внешне напоминает морскую свинку с похожей симпатичной мордочкой, небольшими ушками и большим носом.

В переводе с языка индейцев гуарани «капибара» — это «господин трав». В странах Южной и Центральной Америки это животное называют по-разному — корлинчо, капугиа, капринчо, пончо.

Небольшие глаза находятся высоко на голове, несколько сзади. Рудиментарный хвост. Довольно короткие конечности. Толстая верхняя губа, округлые, короткие уши, широко расставленные ноздри. Задние лапы капибары имеют по три пальца, передние — по четыре, причем между пальцами у нее, как у множества водоплавающих млекопитающих перепонки.

$$\text{КПД} = 1/3$$
[illegible]

Контент vs информация

1. Контент - текст + изображения + видео + другие данные на странице (в т.ч. стили)
2. Информация - семантический уровень данных(смысл)

Мы умеем работать только с контентом

Полезный контент - подмножество всего контента на странице. Данные, полезные для индексации и поиска.

Постановка проблемы (идеальный мир)

Полезный контент идёт в индекс

Больше **разнообразного** полезного контента - больше полнота индекса

Цель: *качать* больше разнообразного контента

Постановка проблемы (реальный мир)

Мы не можем заранее сказать, какой контент находится на странице

Можем только предполагать

Цель 1: качать меньше потенциальных дубликатов

Цель 2: не допускать попадание дубликатов в индекс => поиск дубликатов *после* выкачки

Какие бывают дубликаты?

- Зеркала - совпадение 85-100% **всего контента**
- Плагиат - совпадение 85-100% **полезного контента**

Плагиат

Нашли ошибку?
ctrl+enter

report to orphus

Сайт подключен к системе Orphus. Если Вы увидели ошибку и хотите, чтобы она была устранена, выделите соответствующий фрагмент текста и нажмите Ctrl+Enter.



ANNALES
DE
L'ORDRE
DE
L'ÉTOILE

[Назад К содержанию](#) [Далше](#)

[Разновидности турниров]

I. «Механический» реннен

(нем. Geschiftrennen)

Всадник одет в реннцойг, под доспехом — толстая ватная куртка — *вамс* с рукавами-буфами на упругой подкладке, заменяющими наручи. Ноги зачастую не имеют поножей. Защитой бедра служат ребристые *набедерные щитки* (нем. Streiftartschen, рис. 621) или *дильже* (рис. 622) на ремнях, перекинутых или продернутых через седло. *Легкие реннен седла* (ит. silla rasa) не имеют передних и задних луков (рис. 623). Лошадь покрыта кожаной попоной, голова защищена глухим налобником. В этом виде поединков было две разновидности. [405]

1. «Механический» реннен с тарчем

(нем. Geschifttartschenrennen)

При этом виде турнира удачный удар по тарчу противника позволял оторвать его от кирасы вместе со множеством металлических крепежных деталей и выбросить тарч за голову всадника высоко в воздух. Этот эффект был вызван пружинным механизмом, установленным по центру нагрудника кирасы и соединенным с тарчем посредством штыря. Штырь проходил через отверстие в тарче и заклинивался снаружи металлической шайбой. Между тарчем и пружинным механизмом зажаты концентрические клинья таким образом, что они своим давлением на тарч удерживали пружину механизма, который своим усилием прижимал клинья.

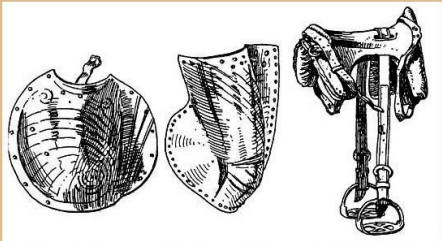


Рис. 621. Набедренный щиток, для защиты бедра от удара о барьер. Кон. XVI в.

Рис. 622. Дильже для правой ноги. Кон. XV в.

Рис. 623. Легкое седло для турнира реннен. Кон. XV в.

★ Запомнить сайт

Словарь на свой сайт

RU ▾

АКАДЕМИК

dic.academic.ru

Словари и энциклопедии на Академике

Введите текст для поиска по словарям и энциклопедиям

Найти!

Энциклопедия средн...

Толкования

Переводы

Книги

Энциклопедия средневекового оружия

Разновидности турниров это:

Толкование

Разновидности турниров

I. «Механический» реннен

(нем. Geschiftrennen)

Всадник одет в реннцойг, под доспехом — толстая ватная куртка — *вамс* с рукавами-буфами на упругой подкладке, заменяющими наручи. Ноги зачастую не имеют поножей. Защитой бедра служат ребристые набедерные щитки (нем. Streiftartschen, рис. 621) или дильже (рис. 622) на ремнях, перекинутых или продернутых через седло. Легкие реннен седла (ит. silla rasa) не имеют передних и задних луков (рис. 623). Лошадь покрыта кожаной попоной, голова защищена глухим налобником. В этом виде поединков было две разновидности.

1. «Механический» реннен с тарчем

(нем. Geschifttartschenrennen)

При этом виде турнира удачный удар по тарчу противника позволял оторвать его от кирасы вместе со множеством металлических крепежных деталей и выбросить тарч за голову всадника высоко в воздух. Этот эффект был вызван пружинным механизмом, установленным по центру нагрудника кирасы и соединенным с тарчем посредством штыря. Штырь проходил через отверстие в тарче и заклинивался снаружи металлической шайбой. Между тарчем и пружинным механизмом зажаты концентрические клинья таким образом, что они своим давлением на тарч удерживали пружину механизма, который своим усилием прижимал клинья.

74

Какие бывают дубликаты?

- Зеркала - совпадение 85-100% **всего контента**
- Плагиат - совпадение 85-100% **полезного контента**

Коды ответов

200 – успех! - их качает спайдер

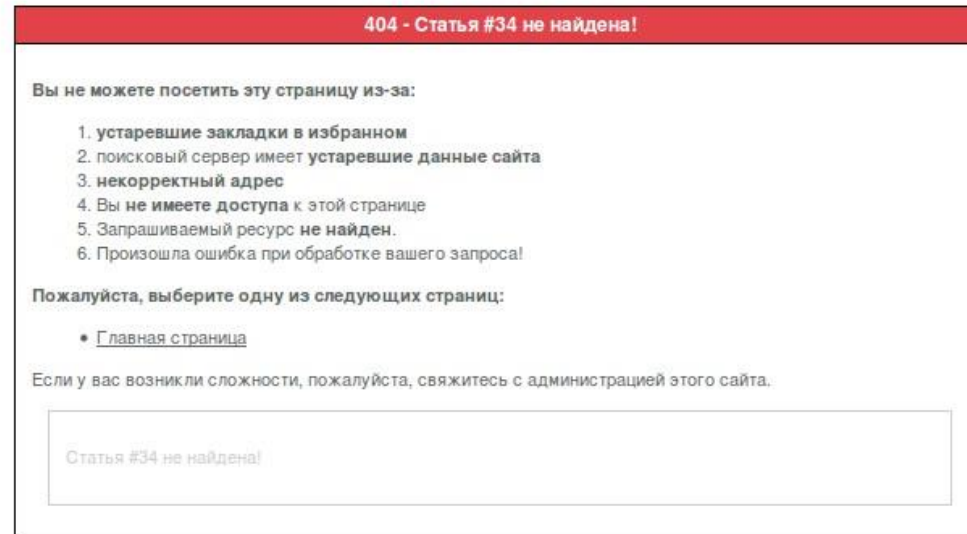
30х – редирект

404 – страница не существует - нет контента для спайдера

50х – ошибка сервера

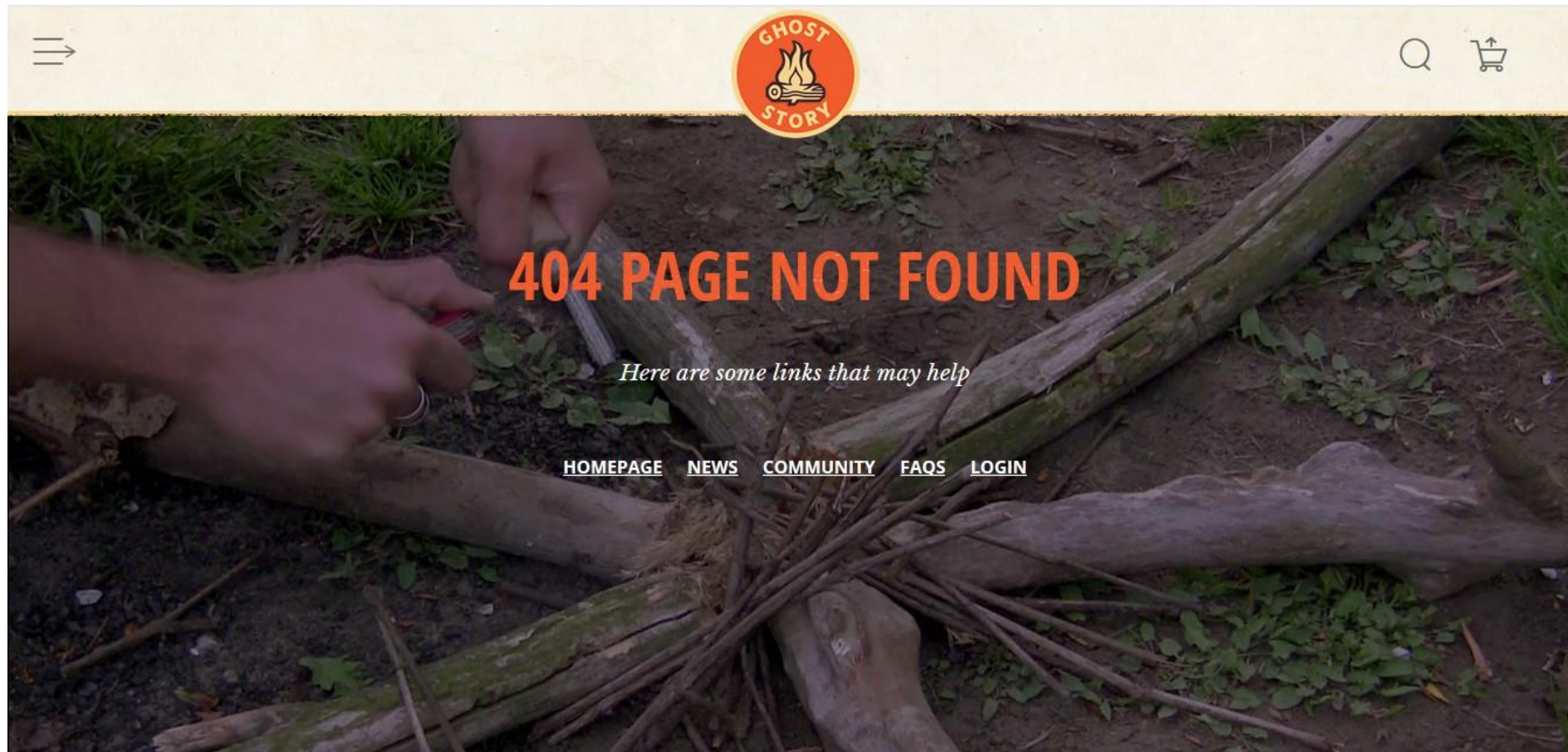
Страница не найдена. Примеры

404: <http://war-toys.ru/component/content/article/34/1-2012-01-28-09-03-06>



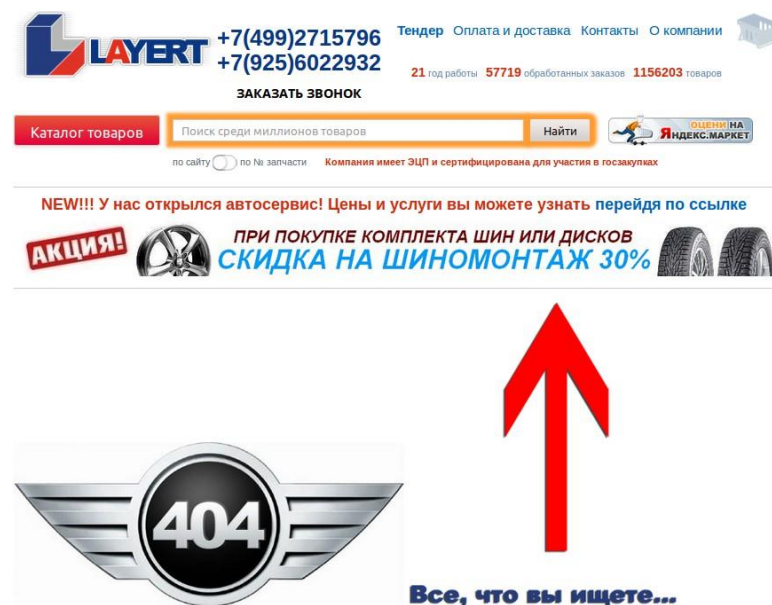
Страница не найдена. Примеры

404: <https://www.ghoststorygames.com/asdfasfsadfs>



Страница не найдена. Примеры

А бывает когда 404я страница отдает 200



Запрошенной страницы не существует.

Это возможно при следующих обстоятельствах:

1. ссылка, по которой вы перешли, устарела
2. вы набрали в адресной строке неверный адрес

Если вы попали на эту страницу по ссылке на нашем сайте, напишите пожалуйста откуда и куда вы хотели попасть на copy@layert.ru

Отсюда вы можете:

1. Вернуться на главную страницу сайта

Виды дубликатов. Soft 404

404

“сайт заблокирован”

“сайта больше нет”

пользователя не существует

и т.д.

Виды дубликатов. Похожие новости

Вечерние пригородные электрички №6095 и №6096 не будут курсировать по маршруту Тайга - Томск-1 - Тайга 7,9 и 15 октября в связи ремонтом на перегоне Богашево - Томск. Об этом сообщает пресс-служба ведомства.

Компания «Кузбасс-пригород» просит пассажиров быть внимательными и планировать свои поездки заранее с учетом изменений в расписании движения пригородных поездов.

Более подробную информацию о расписании движения электричек можно получить в кассах ОАО «Кузбасс-пригород», на сайте компании, а также с 8:00 до 20:00 по телефонам: (3842) 32-37-17, (38448) 7-20-54, 8(905) 968-90-70.

Ранее сообщалось, что РЖД отменит пригородных электричек из Томска и изменят частоту еще одного пригородного поезда из-за перехода на зимнее расписание.

Электропоезда №6095 и №6096 не будут совершать поездки по маршруту Тайга – Томск-1 – Тайга три дня в октябре из-за ремонтных работ, сообщает пресс-служба Западно-Сибирской железной дороги (филиал ОАО «РЖД»).

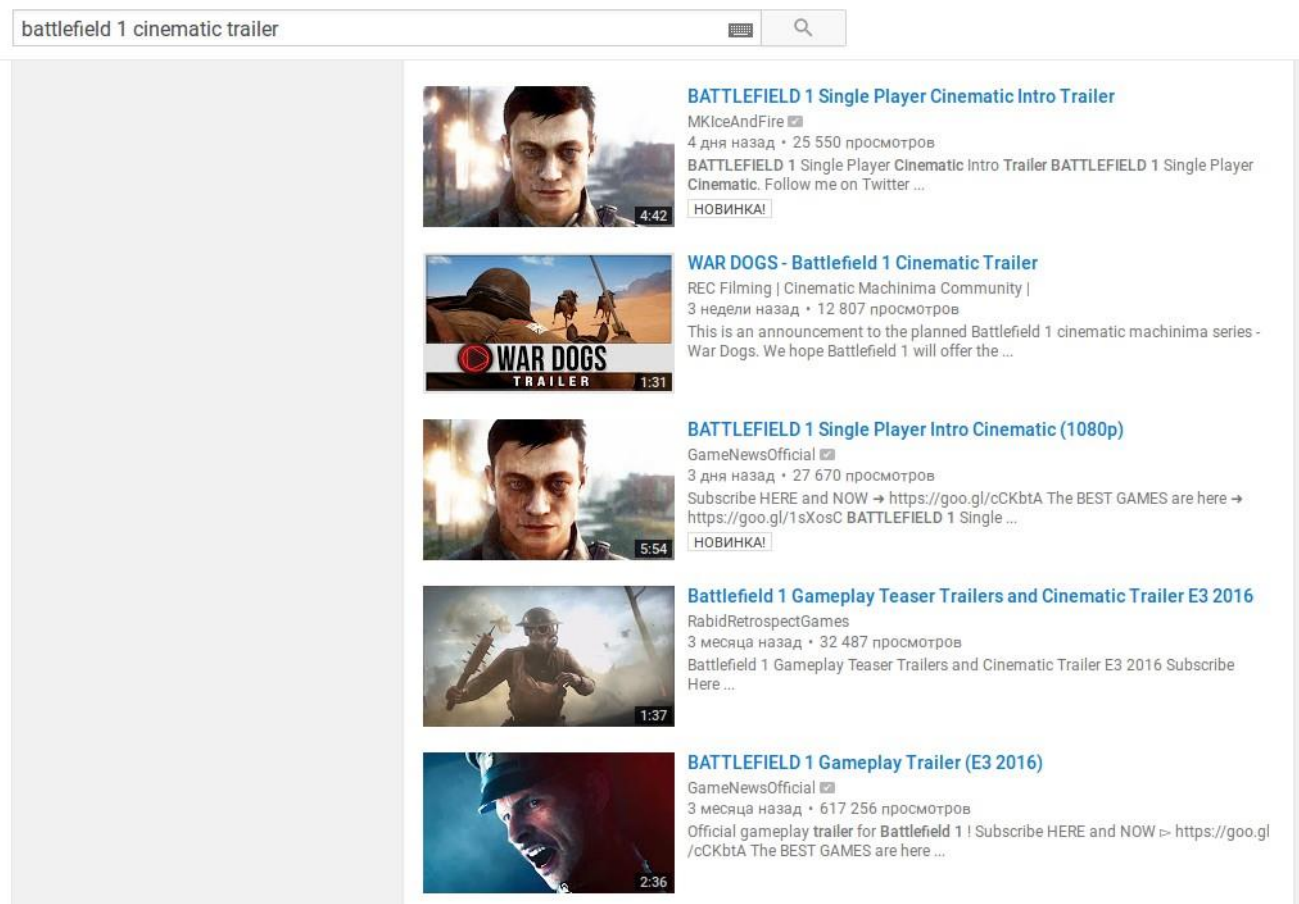
Вечерние пригородные электрички №6095 и №6096 не будут курсировать по маршруту Тайга – Томск-1 – Тайга 7, 9 и 15 октября в связи с проведением капитального ремонта на перегоне Богашево – Томск Кузбасского региона Западно-Сибирской железной дороги.

Компания «Кузбасс-пригород» просит пассажиров планировать свои поездки заранее с учетом изменений в расписании движения пригородных поездов.

Более подробную информацию о расписании движения электричек можно получить в кассах ОАО «Кузбасс-пригород», а также с 08:00 до 20:00 по телефонам 8 (3842) 32-37-17, 8 (3844) 87-20-54.

Виды дубликатов

Дубликатами могут быть не только текстовые документы



Поиск дубликатов

Дано: 2 документа

Задание: определить, являются ли они дубликатами

Поиск дубликатов. Подходы

1. Использовать весь текст
2. Использовать фрагмент текста
3. Использовать несколько фрагментов текста
4. Словари
5. Число/числа, вычисленные на основе особенностей текста
6. Др. сигнатура

Поиск дубликатов. Метрики

Характер сигнатуры определяет допустимое множество метрик

Метрика - функция(!), которая задает отношение между текстами

Поиск дубликатов. Простой пример

Мама мыла раму

vs

Мамма мыла раму

Поиск дубликатов. Шинглы

«Shingle» - «чешуйка», «черепица»

Шинглирование - получение множества фрагментов исходного текста

1 шингл - фрагмент текста длиной N

Поиск дубликатов. Шинглы. Разбиение текста

Мама мыла раму

Как построим шинглы?

Поиск дубликатов. Шинглы. Разбиение текста. Последовательность шинглов

Мама_м_ыла_раму $N = 3$

{"Мам", "а_м", "ыла", "_ра", "му"}

Поиск дубликатов. Шинглы. Разбиение текста. Последовательность шинглов

Мама мыла раму $N = 3$

{"Мам", "а_м", "ыла", "_ра", "му"}

Что делать с группой, меньше чем N ?

Слишком чувствительно к неточным совпадениям:

"мамма мыла раму" -> {"мам", "ма_", "мыл", "а_p", "аму"}

Поиск дубликатов. Шинглы. Разбиение текста. Последовательность шинглов

Мама мыла раму $N = 1$

{"Мама", "мыла", "раму"}

Поиск дубликатов. Шинглы. Разбиение текста. Словарное разбиение

Мама мыла раму $N = 1$

{"Мама", "мыла", "раму"}

Достаточно большие тексты на похожую тематику основываются на практически одинаковых словарях

Иногда порядок важен:

- "Рыцаря нельзя было помиловать, и король решил его казнить"
- "Рыцаря нельзя было казнить, и король решил его помиловать"

Поиск дубликатов. Шинглы. Разбиение текста. Разбиение "внахлѐст"

Мама мыла раму

$N = 10$



Поиск дубликатов. Шинглы. Разбиение текста. Разбиение "внахлѐст"

Мама мыла раму

$N = 10$



shingle2

Поиск дубликатов. Шинглы. Разбиение текста. Разбиение "внахлѐст"

Мама мыла раму

Что делать с конечными шинглами?

$N = 10$



shingle3

Шинглы. Сравнение документов

Построим матрицу смежности:

столбцы - множество документов

строки - всё возможное множество шинглов

	d1	d2	d3	...	dK
sh1	1	1	0		1
sh2	0	1	1		1
sh3	0	1	1		0
...					
shN	1	0	0		1

Шинглы. Сравнение документов

8

Все шинглы длины 8 для [a-zA-Z] -> (26+26+1)

Улучшение - нам не нужно всё множество шинглов. Достаточно множества шинглов из наших документов (т.е. удаляем строки из 0)

Сравнение документов.

У каждого документа – множество шинглов – вектор из 0 и 1. Матрица разреженная, берем только множество синглов документов

	d1	d2	d3	...	dK
sh1	1	1	0		1
sh2	0	1	1		1
sh3	0	1	1		0
...					
shN	1	0	0		1

Сравнение документов. Мера Жаккара

У каждого документа - множество шинглов

Мера Жаккара:

$$JC(A, B) = \frac{A \cap B}{A \cup B}$$

Мера Жаккара. Пример

	d1	d2		
sh1	1	1		
sh2	0	1		
sh3	0	0		
sh4	1	0		
sh5	0	0		
sh6	0	1		

Мера Жаккара. Пример

	d1	d2		
sh1	1	1	*	
sh2	0	1		
sh3	0	0		
sh4	1	0		
sh5	0	0		
sh6	0	1		

Мера Жаккара. Пример

	d1	d2		
sh1	1	1	*	*
sh2	0	1		*
sh3	0	0		
sh4	1	0		*
sh5	0	0		
sh6	0	1		*

Мера Жаккара. Пример

	d1	d2		
sh1	1	1	*	*
sh2	0	1		*
sh3	0	0		
sh4	1	0		*
sh5	0	0		
sh6	0	1		*

$$JC = 1/4$$

Дальнейшие улучшения

- Необходимо уменьшить число попарных сравнений
- Переход от шинглов к числам через хэш-функции
- Сворачиваем к Minshingles для сокращения числа шинглов
- Алгоритм Бродера для сокращения числа попарных сравнений
- LSH

Спасибо за
внимание