



Автоматический сбор данных

Сергей Нурк

Разработчик

Владимир Батыгин

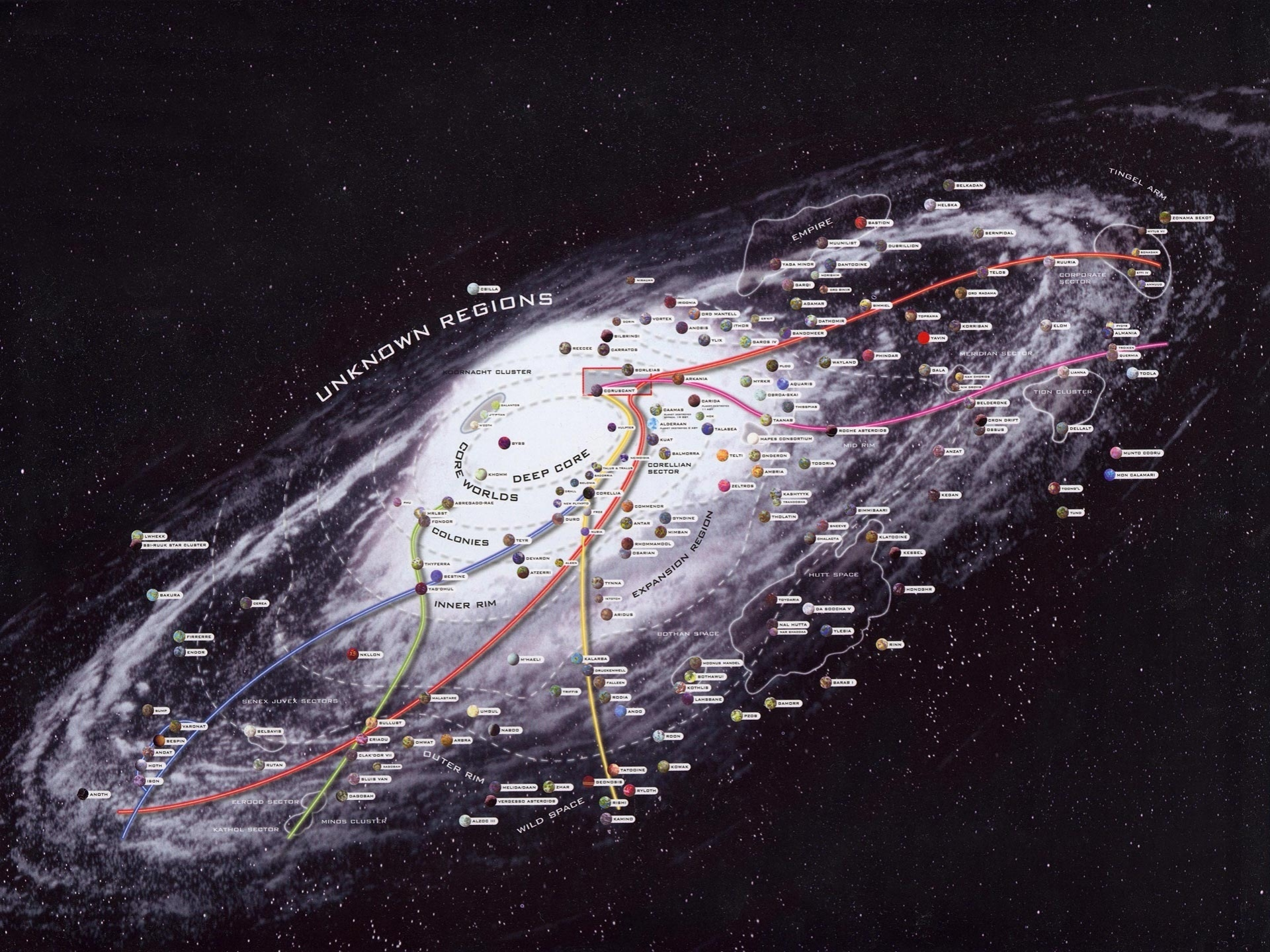
разработчик

АУ, Санкт-Петербург, 1 ноября 2010 года

План

- Введение
- MDR
- SinglePage
- ИТОГИ

Введение



UNKNOWN REGIONS

CORE WORLDS

EXPANSION REGION

WILD SPACE

CORONAUGHT CLUSTER

DEEP CORE

COLONIES

INNER RIM

OUTER RIM

CORRELIAN SECTOR

BOTHAN SPACE

EMPIRE

CORPORAATE SECTOR

MERIDIAN SECTOR

TION CLUSTER

HUTT SPACE

GENEX JUVEK SECTORS

ELRWOOD SECTOR

KATHOL SECTOR

MINDOS CLUSTER

TINGEL ARM

ZONAMA BEXTU

UNHEEK

BAKURA

FIRREERE

ENDOR

YAROHAT

ANGAR

HOTH

TECH

DABOBAN

MELIDA-DAAN

ALZED III

RULLUST

ERIAOU

OHWAT

ARBA

ABAND

SEONDOR

RYLOTH

BIESH

KAMINO

MELSET

FONDOR

YAD'OHUL

BESTINE

ATZERU

TYNNA

ARIDUS

M'HAELI

KALARRA

TRUKKEMELL

FALLER

TRIPPA

RODIA

ANDD

UNHEEK

BAKURA

FIRREERE

ENDOR

YAROHAT

ANGAR

HOTH

TECH

DABOBAN

MELIDA-DAAN

ALZED III

RULLUST

ERIAOU

OHWAT

ARBA

ABAND

SEONDOR

RYLOTH

UNHEEK

BAKURA

FIRREERE

ENDOR

YAROHAT

ANGAR

HOTH

TECH

DABOBAN

MELIDA-DAAN

ALZED III

RULLUST

ERIAOU

OHWAT

ARBA

ABAND

SEONDOR

RYLOTH

CORONAUGHT CLUSTER

DEEP CORE

COLONIES

INNER RIM

OUTER RIM

CORRELIAN SECTOR

BOTHAN SPACE

EMPIRE

CORPORAATE SECTOR

MERIDIAN SECTOR

TION CLUSTER

HUTT SPACE

GENEX JUVEK SECTORS

ELRWOOD SECTOR

KATHOL SECTOR

MINDOS CLUSTER

TINGEL ARM

ZONAMA BEXTU

UNHEEK

BAKURA

FIRREERE

ENDOR

YAROHAT

ANGAR

HOTH

TECH

DABOBAN

MELIDA-DAAN

ALZED III

RULLUST

ERIAOU

OHWAT

ARBA

ABAND

SEONDOR

RYLOTH

BIESH

KAMINO

CORONAUGHT CLUSTER

DEEP CORE

COLONIES

INNER RIM

OUTER RIM

CORRELIAN SECTOR

BOTHAN SPACE

EMPIRE

CORPORAATE SECTOR

MERIDIAN SECTOR

TION CLUSTER

HUTT SPACE

GENEX JUVEK SECTORS

ELRWOOD SECTOR

KATHOL SECTOR

MINDOS CLUSTER

TINGEL ARM

ZONAMA BEXTU

UNHEEK

BAKURA

FIRREERE

ENDOR

YAROHAT

ANGAR

HOTH

TECH

DABOBAN

MELIDA-DAAN

ALZED III

RULLUST

ERIAOU

OHWAT

ARBA

ABAND

SEONDOR

RYLOTH

BIESH

KAMINO

CORONAUGHT CLUSTER

DEEP CORE

COLONIES

INNER RIM

OUTER RIM

CORRELIAN SECTOR

BOTHAN SPACE

EMPIRE

CORPORAATE SECTOR

MERIDIAN SECTOR

TION CLUSTER

HUTT SPACE

GENEX JUVEK SECTORS

ELRWOOD SECTOR

KATHOL SECTOR

MINDOS CLUSTER

TINGEL ARM

ZONAMA BEXTU

UNHEEK

BAKURA

FIRREERE

ENDOR

YAROHAT

ANGAR

HOTH

TECH

DABOBAN

MELIDA-DAAN

ALZED III

RULLUST

ERIAOU

OHWAT

ARBA

ABAND

SEONDOR

RYLOTH

BIESH

KAMINO

CORONAUGHT CLUSTER

DEEP CORE

COLONIES

INNER RIM

OUTER RIM

CORRELIAN SECTOR

BOTHAN SPACE

EMPIRE

CORPORAATE SECTOR

MERIDIAN SECTOR

TION CLUSTER

HUTT SPACE

GENEX JUVEK SECTORS

ELRWOOD SECTOR

KATHOL SECTOR

MINDOS CLUSTER

TINGEL ARM

ZONAMA BEXTU

UNHEEK

BAKURA

FIRREERE

ENDOR

YAROHAT

ANGAR

HOTH

TECH

DABOBAN

MELIDA-DAAN

ALZED III

RULLUST

ERIAOU

OHWAT

ARBA

ABAND

SEONDOR

RYLOTH

BIESH

KAMINO

CORONAUGHT CLUSTER

DEEP CORE

COLONIES

INNER RIM

OUTER RIM

CORRELIAN SECTOR

BOTHAN SPACE

EMPIRE

CORPORAATE SECTOR

MERIDIAN SECTOR

TION CLUSTER

HUTT SPACE

GENEX JUVEK SECTORS

ELRWOOD SECTOR

KATHOL SECTOR

MINDOS CLUSTER

TINGEL ARM

ZONAMA BEXTU

UNHEEK

BAKURA

FIRREERE

ENDOR

YAROHAT

ANGAR

HOTH

TECH

DABOBAN

MELIDA-DAAN

ALZED III

RULLUST

ERIAOU

OHWAT

ARBA

ABAND

SEONDOR

RYLOTH

BIESH

KAMINO

CORONAUGHT CLUSTER

DEEP CORE

COLONIES

INNER RIM

OUTER RIM

CORRELIAN SECTOR

BOTHAN SPACE

EMPIRE

CORPORAATE SECTOR

MERIDIAN SECTOR

TION CLUSTER

HUTT SPACE

GENEX JUVEK SECTORS

ELRWOOD SECTOR

KATHOL SECTOR

MINDOS CLUSTER

TINGEL ARM

ZONAMA BEXTU

UNHEEK

BAKURA

FIRREERE

ENDOR

YAROHAT

ANGAR

HOTH

TECH

DABOBAN

MELIDA-DAAN

ALZED III

RULLUST

ERIAOU

OHWAT

ARBA

ABAND

SEONDOR

RYLOTH

BIESH

KAMINO

CORONAUGHT CLUSTER

DEEP CORE

COLONIES

Databank:

By Category

By Appearance

Millennium Falcon



From the Movies

A legendary starship despite its humble origins and deceptively dilapidated exterior, the *Millennium Falcon* has factored into some of the Rebel Alliance's greatest victories over the Empire. On the surface, the *Falcon* looks like any other

Corellian freighter, with a saucer-shaped primary hull, a pair of forward cargo-gripping mandibles, and a cylindrical cockpit mounted to the ship's side.

Beneath its hull, though, the *Falcon* packs many powerful secrets. Its owners made "special modifications" on the freighter, boosting its speed, shielding and performance to downright illegal levels. Its weaponry has been upgraded to military-class quad laser turrets. To cover rapid escapes, the *Falcon* sports a ventrally mounted hatch-concealed antipersonnel repeating laser. Between its forward mandibles rest concussion missile launchers. The habitable interior of the vessel also has a few surprises, such as concealed scanner-proof smuggling compartments.

The *Falcon* pays a heavy price for its augmented performance, though. It is extremely recalcitrant and often unpredictable. Its reconditioned hyperdrive often fails. Its current captain, Han

STAR WARS profile

Appeared in:

I II III IV V VI CW

Homeworld:

Corellia

Size:

26.7 meters long

Weapon:

quad laser cannons, concussion missiles

Affiliation:

Smuggling, Rebel Alliance

Type:

YT-1300 freighter

Manufacturer:

Related News

Celebration V: Slave Leia 360

360° of Leia's Metal Bikini at Celebration VI Check out the second 360 view on the *Millennium Falcon* set!

Eat Lunch at Lightspeed: *Star Wars* Sandwich Cutters

Have lunch with Darth Vader and Han Solo with TIE fighter and *Millennium Falcon* sandwich cutters from Williams-Sonoma!

"This is No Cave" Exclusive Print Now Available
The fifth in *The Empire Strikes Back* 30 Anniversary Artwork Series is now available exclusively at StarWarsShop and limited to just 100 pieces!

STAR WARS SHOP

more product, more exclusives



The Sounds of *Star Wars*
Our Price: \$ 59.99



Star Wars 3D Lenticular Playing Cards Tin
Our Price: \$ 24.99

Millennium Falcon



STAR WARS profile

Appeared in:

I II III IV V VI CW

Homeworld:

Corellia

Size:

26.7 meters long

Weapon:

quad laser cannons,
concussion missiles

Name

Millennium Falcon

Homeworld

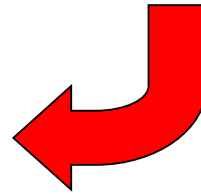
Corellia

Size

26.7 m

Weapon

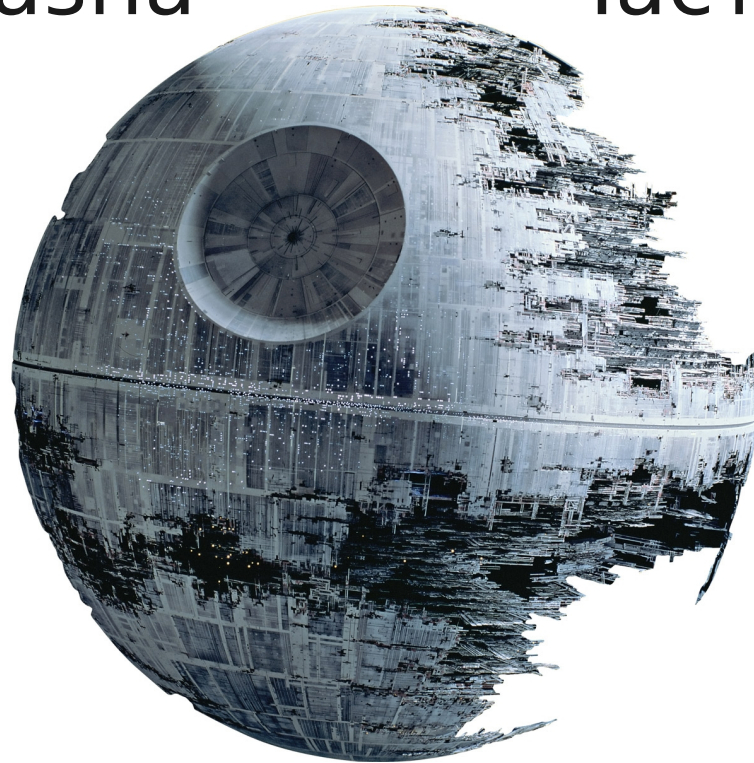
- quad laser cannons
- concussion missiles



Вёрстка

Разнообразна

Часто изменяется



Специализированные парсеры

На каждый сайт – свой



Нужна армия



Общие алгоритмы



Предполагают наличие на странице регулярной структуры

<u>KIT NAME</u>	<u>MANUFACTURER</u>	<u>SCALE</u>	<u>MATERIAL</u>
<u>A-Wing Fighter</u>	SMT	1.48	Resin
<u>A-Wing Fighter</u>	AMT/ERTL	1.48	Plastic
<u>Aayla Secura</u>	Nut-Hut Productions	1.6	Resin
<u>Admiral Ackbar</u>	Craft Master	1.14	Vinyl
<u>Anakin Skywalker</u>	Mojoresin	1.8	Resin
<u>Anakin's Podracer</u>	AMT/ERTL	1.32	Plastic
<u>AT-AT</u>	SMT	1.53	Resin
<u>AT-AT</u>	MPC/ERTL	1.100	Plastic



MDR

Kit name
 A-Wing Fighter
 Manufacturer
 AMT/ERTL
 Scale
 1.48
 Material
 Plastic

Основные требования

- Простая настройка
- Простая поддержка
- Высокие показатели полноты и точности

Глава 1.

MDR



You Have Selected:**Brand:**

Star Wars

Product Category:

Vehicles

X removes selection

Narrow Your Results**Brand**

SpeedStars (1)

Transformers (1)

Product Category

Cars (1)

Remote Controlled (4)

Spaceships (10)

Gender

Boys (16)

Both (1)

Price

Under \$10 (1)

Over \$100 (1)

\$10 - \$20 (4)

\$20 - \$30 (6)

\$30 - \$40 (2)

\$40 - \$50 (1)

\$50 - \$100 (2)

More Ways to Shop

Sort By: - select -

17 Items Found

Items Per Page: 10, 50

1 2 ...Next >



More Images

Star Wars SPEED STARS Millennium Falcon Remote Control
Item #: 19647

The most famous ship in the galaxy flies into action! The Millennium Falcon soars through space, taking the fight against the Empire everywhere it goes. This powerful starship may not look like much, but she's got it where it counts! Get ready for a battle of epic proportions! Featuring.....More

Approx. Retail: \$49.99

Where to Buy STAR WARS



More Images

Star Wars The Clone Wars MagnaGuard Fighter
Item #: 87969

IG-series 100 MagnaGuards – the droid bodyguards for General Grievous – fly specialized fighters as part of their protective duties for the cyborg general. The agile fighters respond to the demands of their droid pilots and are armed with missiles to repel attacks from Republic forces. Sleek.....More

Approx. Retail: \$24.99

Where to Buy STAR WARS



More Images

STAR WARS MILLENNIUM FALCON
Item #: 87591

!!WARNING: CHOKING HAZARD-Small Parts. Not For Children Under 3 Years. Best known as the fastest ship in the Star Wars galaxy, piloted by Han Solo and Chewbacca, this detailed replica of the rebel spacecraft packs powerful secrets and special modifications - inside and out. Measuring more than.....More

Approx. Retail: \$179.99



More Ways to Buy STAR WARS



More Images

Star Wars The Clone Wars Radio Control Hailfire Droid
Item #: 93974

Clone troopers beware - the hailfire droid is rolling into battle, packed with missiles and ready to attack! It cuts a fearsome path as it moves over the ground, blasting anything in its way. Watch out - this mobile battle droid easily changes direction to keep the clone trooper army on the run!More

Approx. Retail: \$59.99



More Ways to Buy STAR WARS



More Images

Star Wars The Clone Wars Republic V-19 Torrent Starfighter
Item #: 19696

After making its debut at the Battle of Geonosis, the Republic's V-19 Torrent assault fighter is used with effectiveness throughout the Clone Wars. Jedi commanders lead clone-piloted V-19 Torrents into battle against vulture droids and Separatist forces. Launch yourself into action with this.....More

Approx. Retail: \$24.99

Where to Buy STAR WARS

Что надо получить

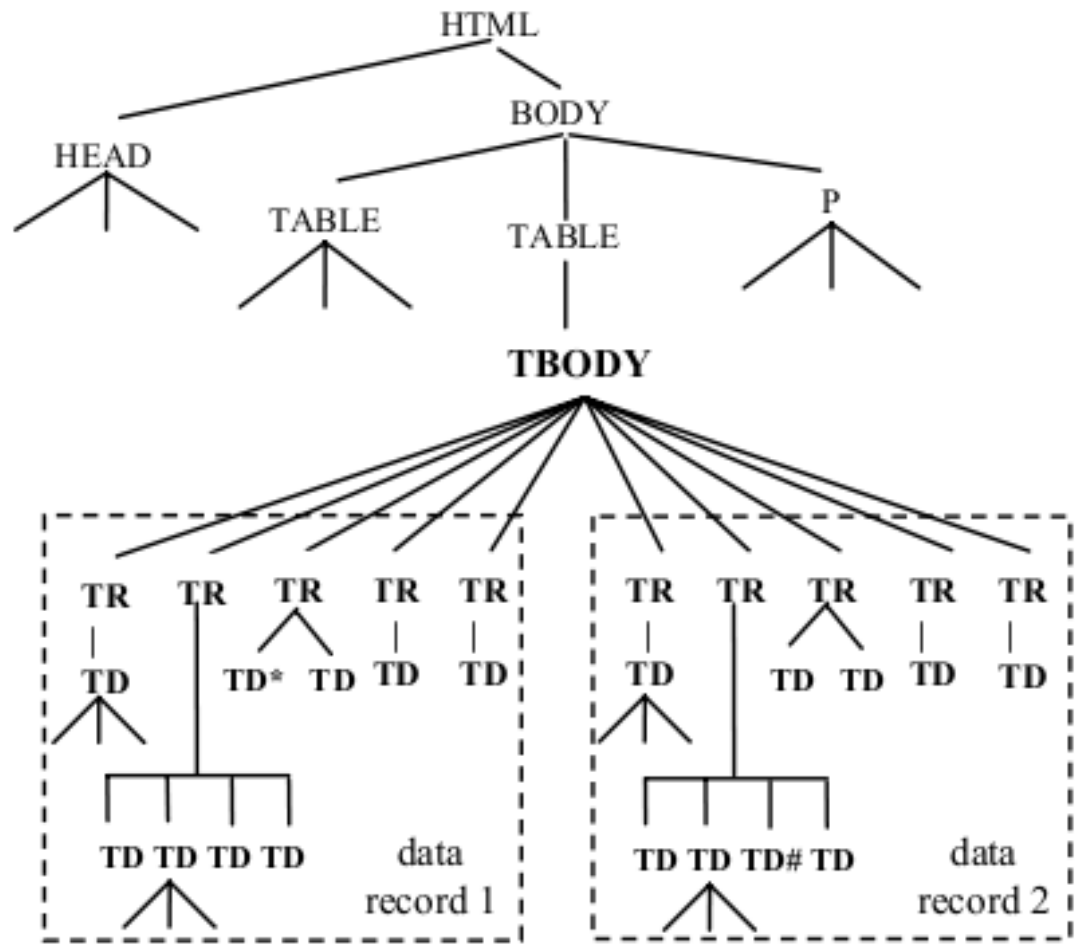
- MILLENNIUM FALCON
- Approx. Retail: \$179.99
- Item #: 87591
- Best known as the fastest ship in the Star Wars galaxy, piloted by Han Solo and Chewbacca ...



Исходный код страницы

```
<html><head> ... </head>
  <body>
    <table> ... </table>
    <table>
      <tr><td>MILLENNIUM FALCON</td></tr>
      <tr>
        <td>Approx. Retail: $49.99</td>
        <td>Item #: 19647</td>
        <td>The most famous ship ...</td>
      </tr>
      ...
      <tr><td>MagnaGuard Fighter</td></tr>
      <tr>
        <td>Approx. Retail: $24.99</td>
        <td>Item #: 87969</td>
        <td>IG-series 100 ...</td>
      </tr>
      ...
```





You Have Selected:**Brand:**

Star Wars

Product Category:

Vehicles

[X] removes selection

Narrow Your Results**Brand**

SpeedStars (1)

Transformers (1)

Product Category

Cars (1)

Remote Controlled (4)

Spaceships (10)

Gender

Boys (16)

Both (1)

Price

Under \$10 (1)

Over \$100 (1)

\$10 - \$20 (4)

\$20 - \$30 (6)

\$30 - \$40 (2)

\$40 - \$50 (1)

\$50 - \$100 (2)

More Ways to Shop

Sort By: - select -

17 Items Found

Items Per Page: 10, 50

1 2 ...Next >

[More Images](#)**Star Wars SPEED STARS Millennium Falcon Remote Control**
Item #: 19647

The most famous ship in the galaxy flies into action! The Millennium Falcon soars through space, taking the fight against the Empire everywhere it goes. This powerful starship may not look like much, but she's got it where it counts! Get ready for a battle of epic proportions! Featuring.....More

Approx. Retail: \$49.99

[Where to Buy STAR WARS](#)[More Images](#)**Star Wars The Clone Wars MagnaGuard Fighter**
Item #: 87969

IG-series 100 MagnaGuards – the droid bodyguards for General Grievous – fly specialized fighters as part of their protective duties for the cyborg general. The agile fighters respond to the demands of their droid pilots and are armed with missiles to repel attacks from Republic forces. Sleek.....More

Approx. Retail: \$24.99

[Where to Buy STAR WARS](#)[More Images](#)**STAR WARS MILLENNIUM FALCON**
Item #: 87591

!!WARNING: CHOKING HAZARD-Small Parts. Not For Children Under 3 Years. Best known as the fastest ship in the Star Wars galaxy, piloted by Han Solo and Chewbacca, this detailed replica of the rebel spacecraft packs powerful secrets and special modifications - inside and out. Measuring more than.....More

Approx. Retail: \$179.99

[BUY NOW](#)[More Ways to Buy STAR WARS](#)[More Images](#)**Star Wars The Clone Wars Radio Control Hailfire Droid**
Item #: 93974

Clone troopers beware - the hailfire droid is rolling into battle, packed with missiles and ready to attack! It cuts a fearsome path as it moves over the ground, blasting anything in its way. Watch out - this mobile battle droid easily changes direction to keep the clone trooper army on the run!More

Approx. Retail: \$59.99

[BUY NOW](#)[More Ways to Buy STAR WARS](#)[More Images](#)**Star Wars The Clone Wars Republic V-19 Torrent Starfighter**
Item #: 19696

After making its debut at the Battle of Geonosis, the Republic's V-19 Torrent assault fighter is used with effectiveness throughout the Clone Wars. Jedi commanders lead clone-piloted V-19 Torrents into battle against vulture droids and Separatist forces. Launch yourself into action with this.....More

Approx. Retail: \$24.99

[Where to Buy STAR WARS](#)

Эвристики

- Основные теги : table, tr, td, div, p, li ...

Эвристики

- Основные теги : table, tr, td, div, p, li ...
- Дисперсия. Считается количество тегов в сущности

Эвристики

- Основные теги : table, tr, td, div, p, li ...
- Дисперсия. Считается количество тегов в сущности
- Повторяющиеся паттерны

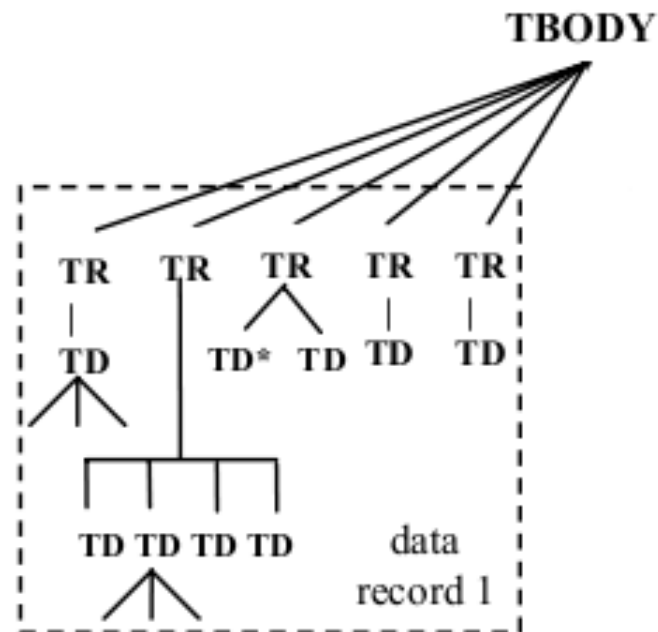
Эвристики

- Основные теги : table, tr, td, div, p, li ...
- Дисперсия. Считается количество тегов в сущности
- Повторяющиеся паттерны
- Доменно-специфичные эвристики



Сущность

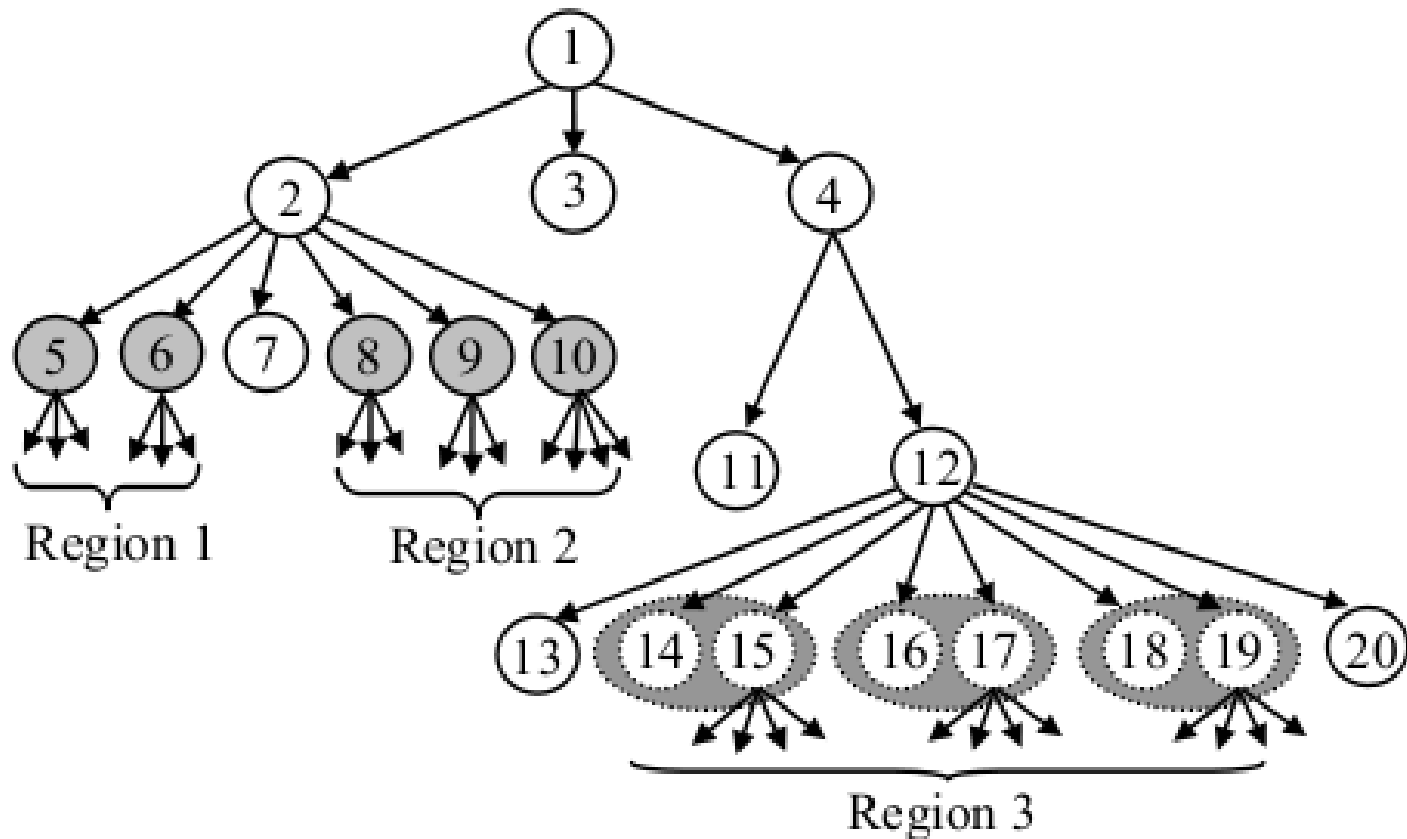
- Все узлы имеют общего родителя
- Все узлы смежны



Области данных

- 2 и более сущности
- Сущности имеют общего родителя
- Сущности имеют одинаковую длину
- Сущности смежны
- Расстояние между сущностями меньше определенного порога

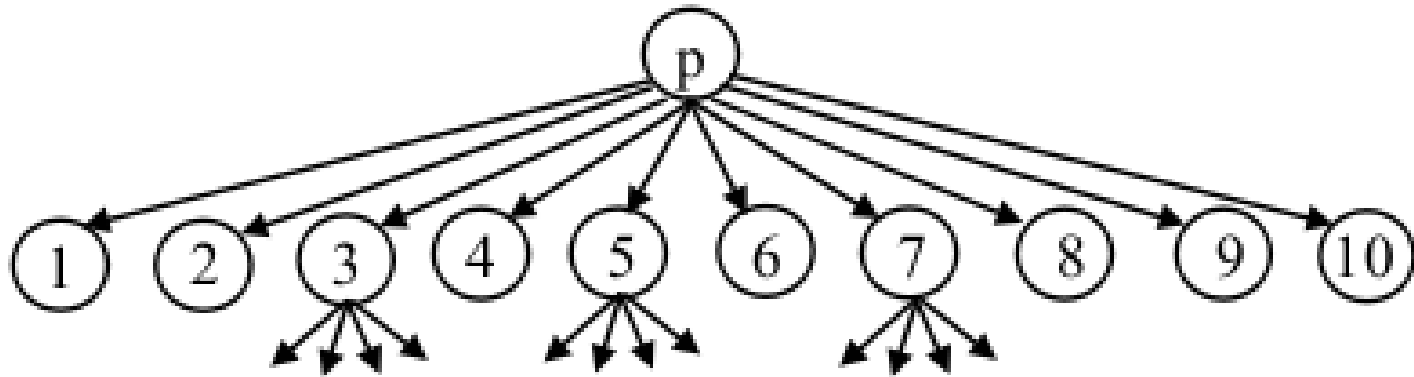
Области данных



Поиск кандидатов

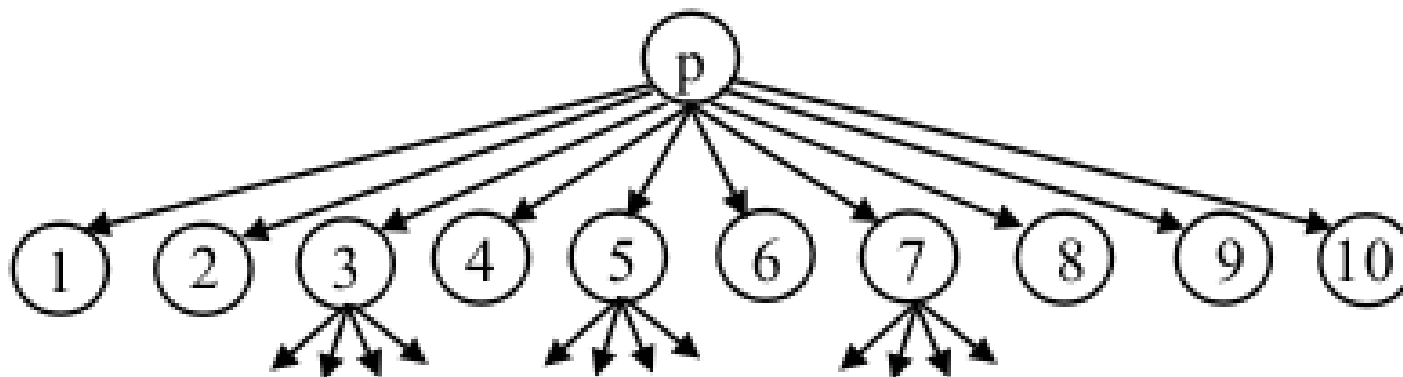
- Спуск в глубину
- На каждом уровне пытаемся найти кандидата
 - Для этого производится подсчет расстояний между предполагаемыми сущностями

Сравнение узлов



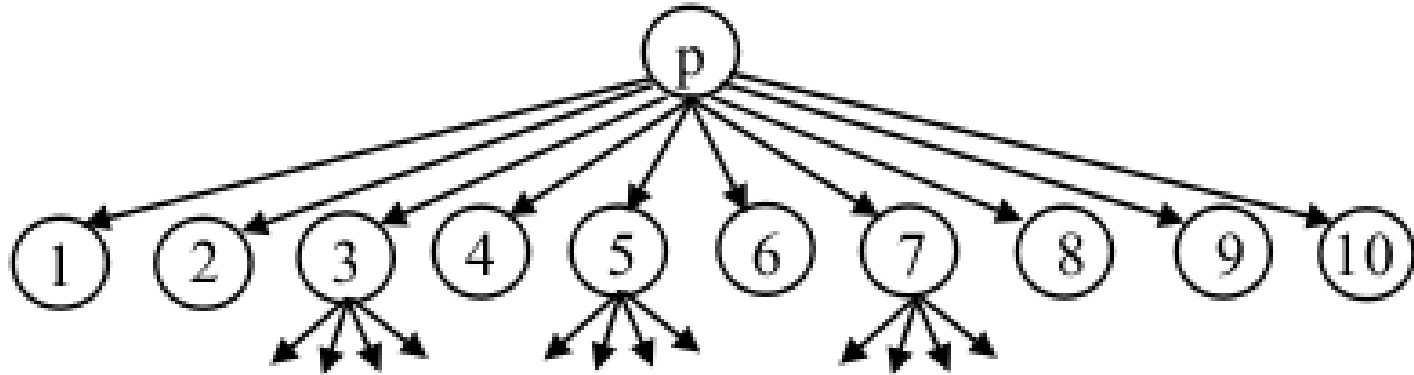
- (1, 2), (2, 3), (3, 4), (4, 5), (5, 6), (6, 7), (7, 8) ...
- (2, 3), (3, 4), (4, 5), (5, 6), (6, 7), (7, 8) ...
- (3, 4), (4, 5), (5, 6), (6, 7), (7, 8) ...
- (4, 5), (5, 6), (6, 7), (7, 8) ...
- (5, 6), (6, 7), (7, 8) ...

Сравнение узлов



- (1, 2), (2, 3), (3, 4), (4, 5), (5, 6), (6, 7), (7, 8), (8, 9), (9, 10)
- (1-2, 3-4), (3-4, 5-6), (5-6, 7-8), (7-8, 9-10)
- (1-2-3, 4-5-6), (4-5-6, 7-8-9)

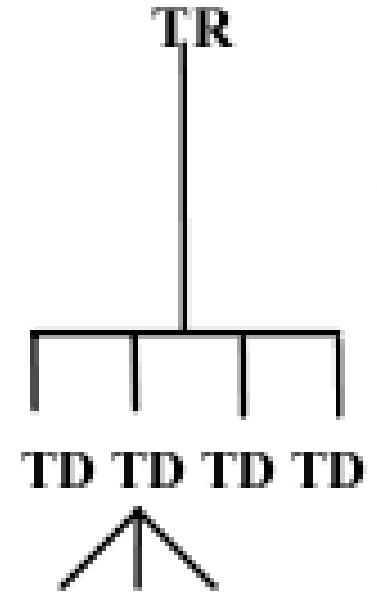
Сравнение узлов



- (1, 2), (8, 9, 10), (3-4, 5-6, 7-8), (2-3, 4-5, 6-7)

Сравнение узлов

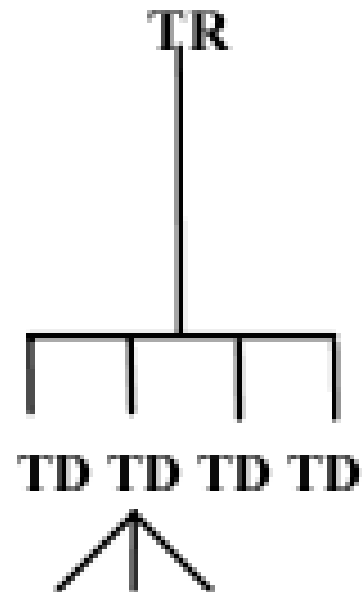
< TR TD TD ... TD TD >



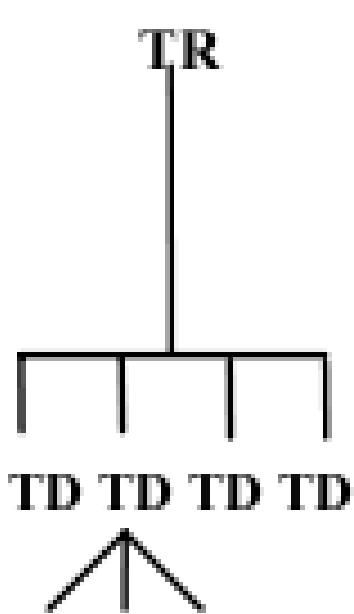
Сравнение узлов

< TR TD TD ... TD TD >

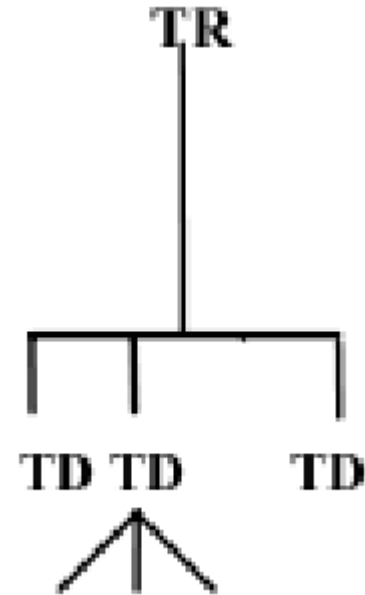
< TR (TD TD (...) TD TD) >



Расстояние Левенштейна



TR TD TD ... TD TD

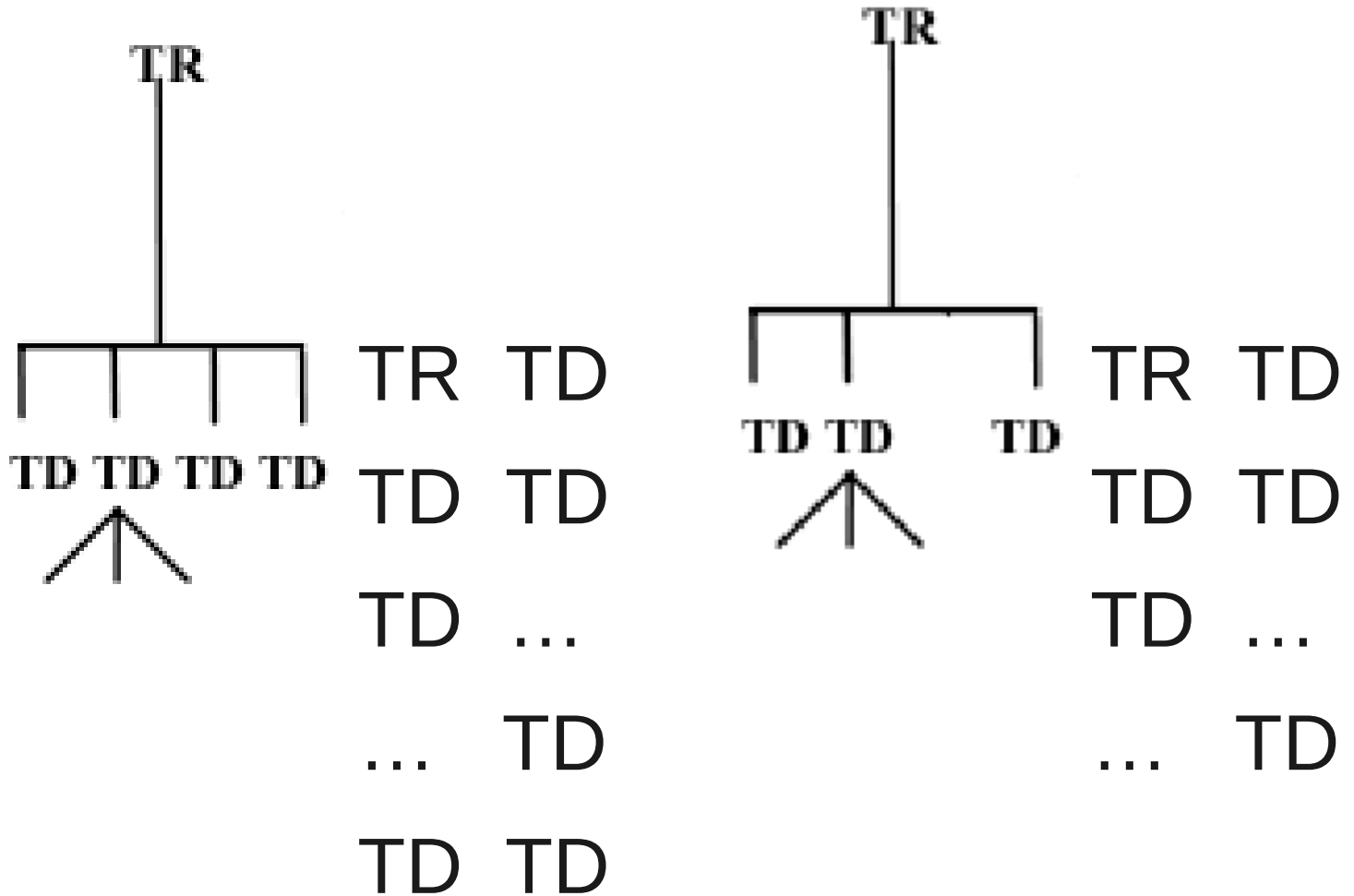


TR TD TD ... TD

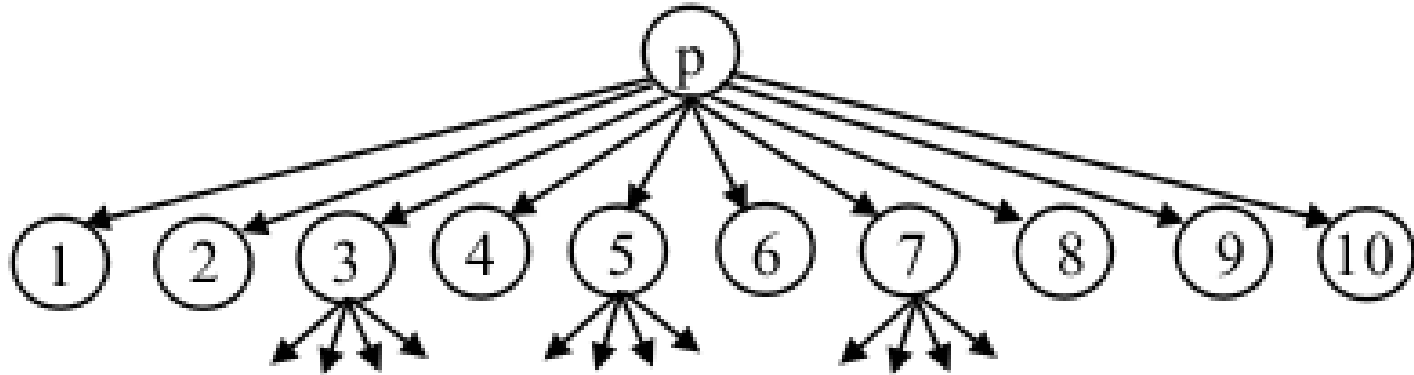
TR TD TD ... TD TD

TR TD TD ... * TD = 1

ШИНГЛЫ



Сравнение узлов

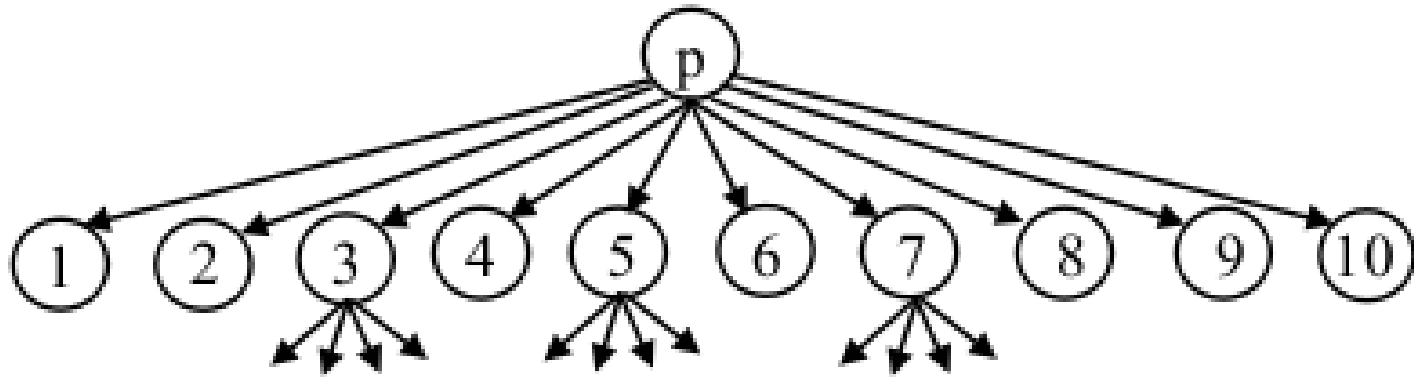


- (1, 2), (8, 9, 10), (3-4, 5-6, 7-8), (2-3, 4-5, 6-7)

Выбор лучшей области данных

- Выбор самой большой
- Выбор самой узкой
- Выбор самой старшей
- Выбор самой первой
- Взвешивание по ключевым словам

Выбор лучшей области данных



- ~~(1, 2), (8, 9, 10), (3-4, 5-6, 7-8), (2-3, 4-5, 6-7)~~

СЛОЖНОСТИ

- Недостаточно регулярная структура

СЛОЖНОСТИ

- Недостаточно регулярная структура
- Мало данных на странице

СЛОЖНОСТИ

- Недостаточно регулярная структура
- Мало данных на странице
- Много шума



Глава 2. SinglePage



Управляемая экстракция

1. Пользователь задает примеры
2. Система автоматически извлекает данные со всего сайта

Millennium Falcon ← name



STAR WARS profile

Appeared in:

I II III IV V VI CW

Homeworld:

Corellia ← made in

Size:

26.7 meters long ← size

Weapon:

quad laser cannons,
concussion missiles ← weapon

Взгляд внутрь

1. По примерам строим шаблоны (один атрибут – один шаблон)
2. Применяем шаблоны к остальным (подходящим) страницам

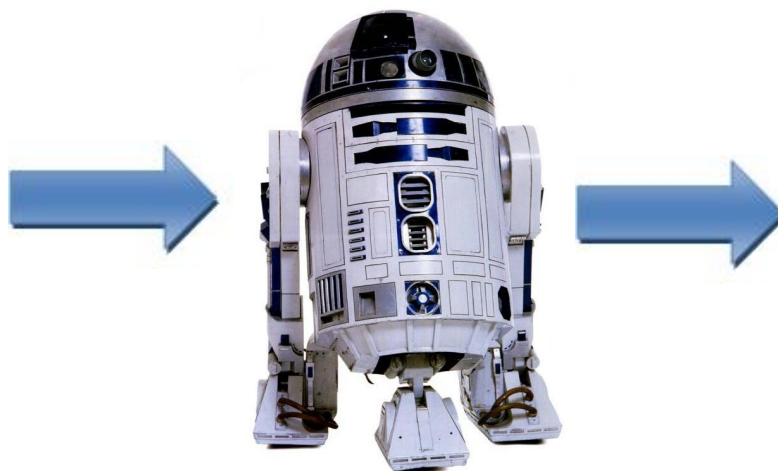
Name
Millennium Falcon

Homeworld
Corellia

Size
26.7 m

Weapon

- quad laser cannons
- concussion missiles




Гипотетические преимущества

- Быстрая настройка
- Не требуется разбираться в структуре страницы
- Структурированное извлечение нужной информации
- Устойчивость к изменениям вёрстки (пересоздание шаблонов)

Ограничения

- Отдельная страница на каждый объект
- Группа страниц с однотипной вёрсткой

Millennium Falcon



STAR WARS profile

Appeared in:
I II III IV V VI CW

Homeworld:
Corellia

Size:
26.7 meters long

Weapon:
quad laser cannons,
concussion missiles

Алгоритм



VS



Millennium Falcon



STAR WARS profile

Appeared in:

I II III IV V VI CW

Homeworld:

Corellia

Size:

26.7 meters long

Weapon:

quad laser cannons,
concussion missiles

```
<ul id="profileContainer">
```

```
<li>Homeworld:<br/><span>Corellia</span></li>
```

```
<li>Size:<br/><span>26.7 meters long</span></li>
```

```
<li>Weapon:<br/><span>quad laser cannons  
, concussion missiles</span></li>
```

STAR WARS profile

Appeared in:

I II III IV V VI CW

Homeworld:

Corellia

Size:

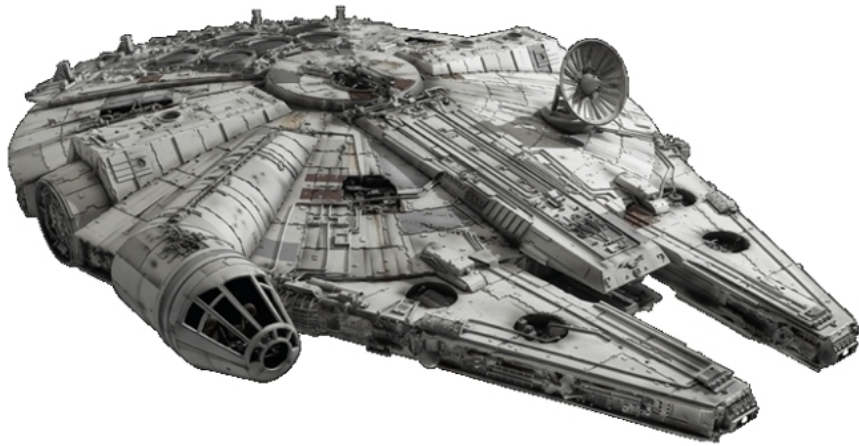
26.7 meters long

Weapon:

quad laser cannons,
concussion missiles

//u/li[3]/span

Millennium Falcon



STAR WARS profile

Appeared in:

I II III IV V VI CW

Homeworld:

Corellia

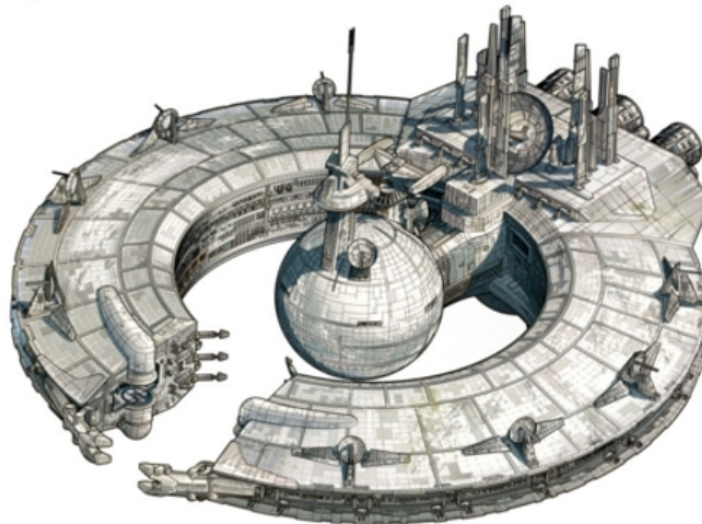
Size:

26.7 meters long

Weapon:

quad laser cannons,
concussion missiles

Droid Control Ship



STAR WARS profile

Appeared in:

I II III IV V VI CW

Size:

3,170 meters diameter

Weapon:

20 quad laser cannons;
4 turbolasers

STAR WARS profile

Appeared in:

I II III IV V VI CW

Homeworld:

Corellia

Size:

26.7 meters long

Weapon:

quad laser cannons,
concussion missiles

STAR WARS profile

Appeared in:

I II III IV V VI CW

Size:

3,170 meters diameter

Weapon:

20 quad laser cannons;
4 turbolasers

//u/li[3]/span

STAR WARS profile

Appeared in:

I II III IV V VI CW

Homeworld:

Corellia

Size:

26.7 meters long

Weapon:

quad laser cannons,
concussion missiles

STAR WARS profile

Appeared in:

I II III IV V VI CW

Size:

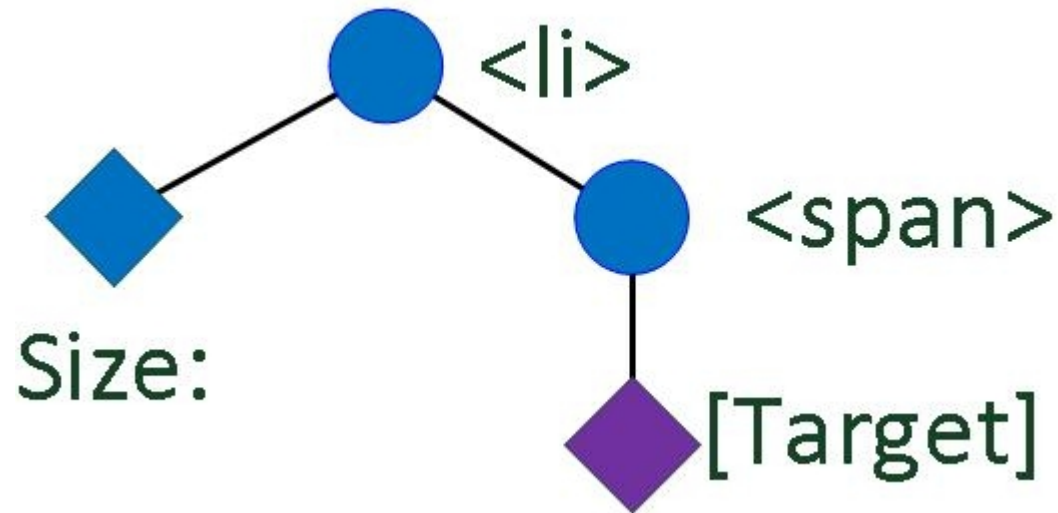
3,170 meters diameter

Weapon:

20 quad laser cannons;
4 turbolasers

//u/li[3]/span  Size=20 quad laser...

TreePattern



```
<ul id="profileContainer">
  <li>Homeworld:<br/><span>Coreellia</span></li>
  <li>&u>Size:<br/><span>&u>26.7 meters long</span></li>
  <li>Weapon:<br/><span>quad laser cannons
    , concussion missiles</span></li>
</ul>
```

STAR WARS profile

Appeared in:

I II III IV V VI CW

Size:

3,170 meters diameter

Weapon:

20 quad laser cannons;
4 turbolasers

TreePattern

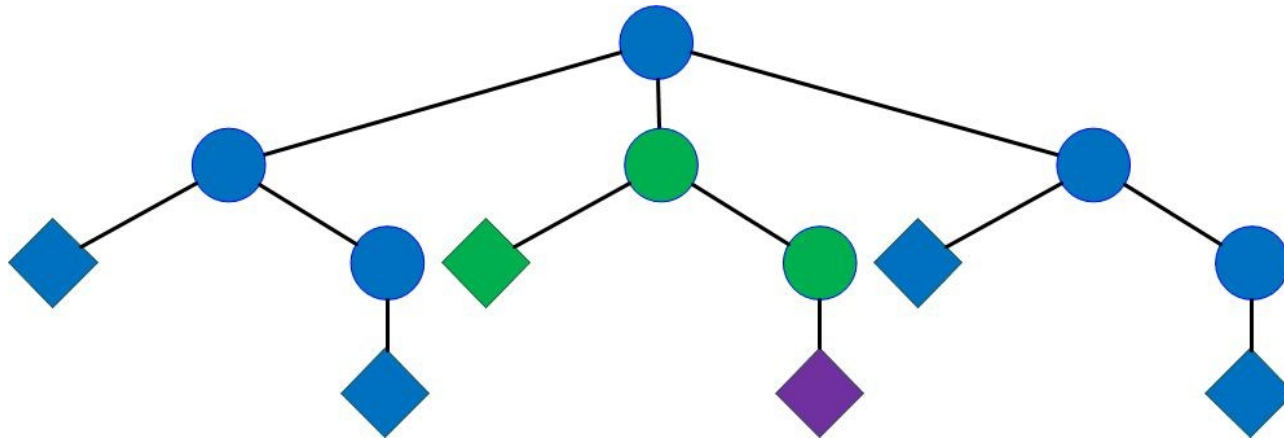
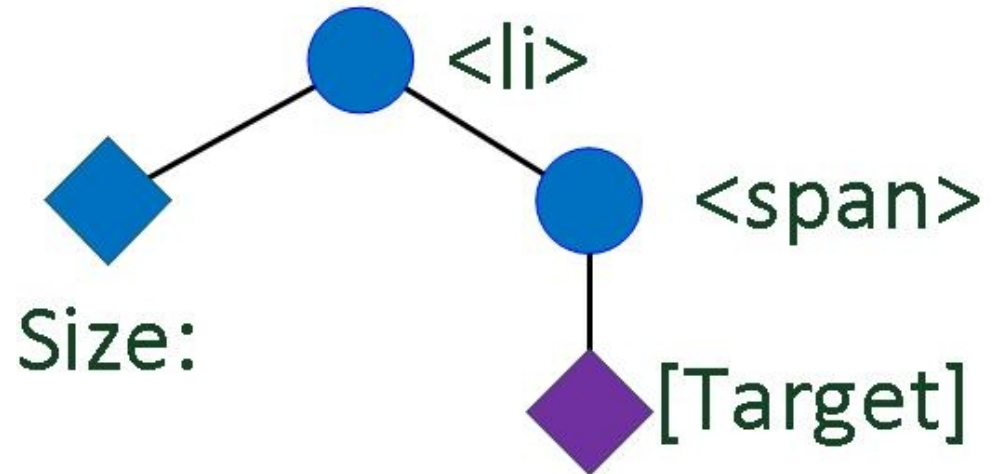


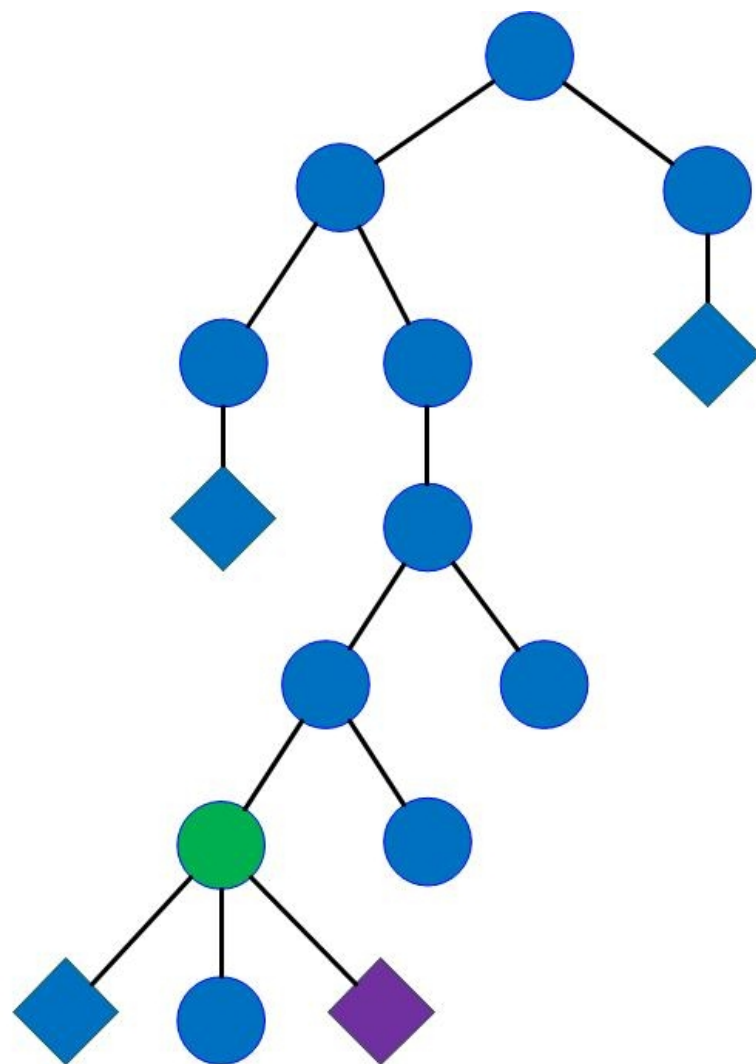
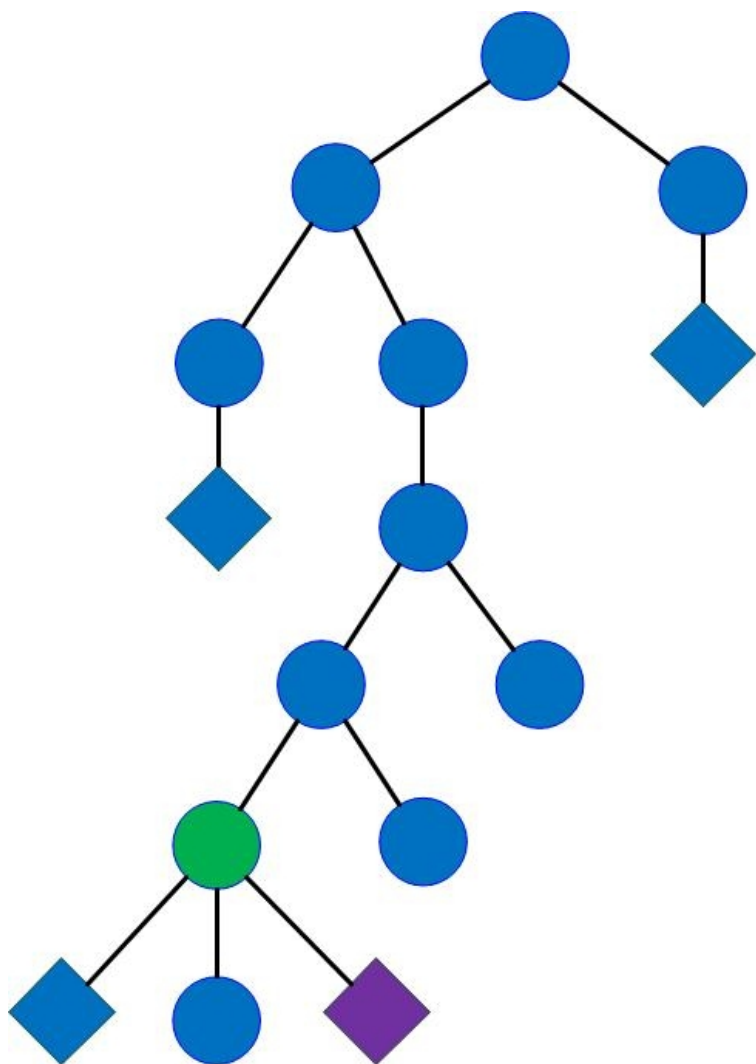
Схема алгоритма

- Находим вхождения примеров
- Строим `TreePattern`, задающий положение узла
- Применяем `TreePattern` к остальным страницам

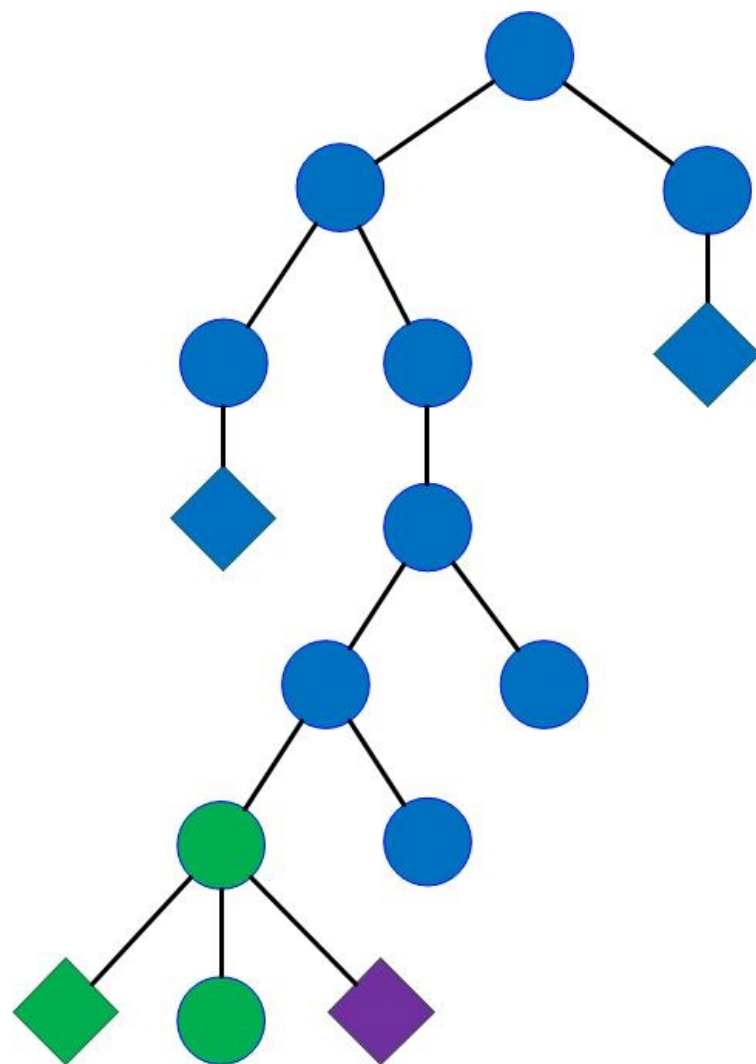
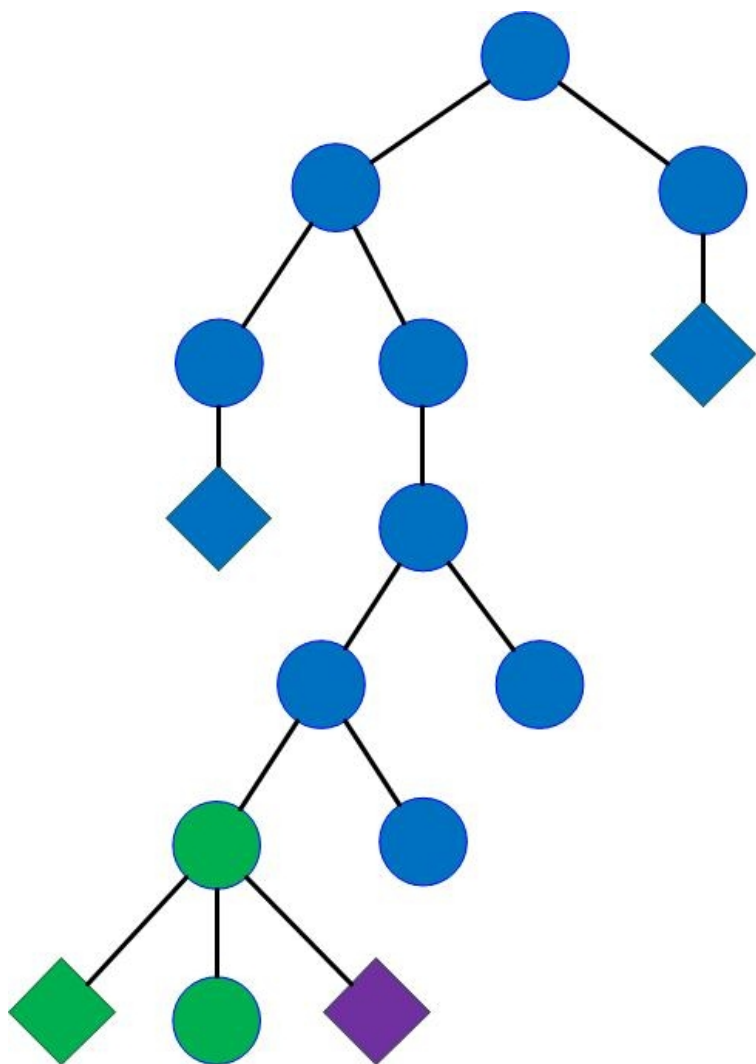
Построение TreePattern

- Начинаем от мест вхождения примеров
- Синхронно движемся по деревьям, сравнивая узлы
- Накапливаем TreePattern

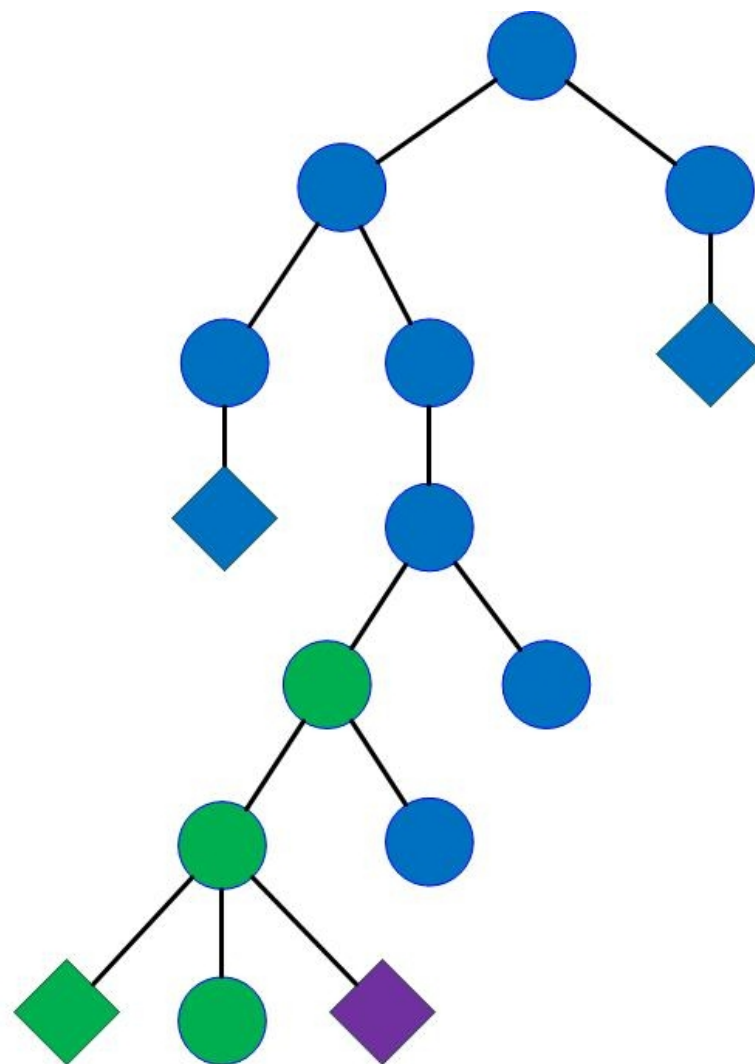
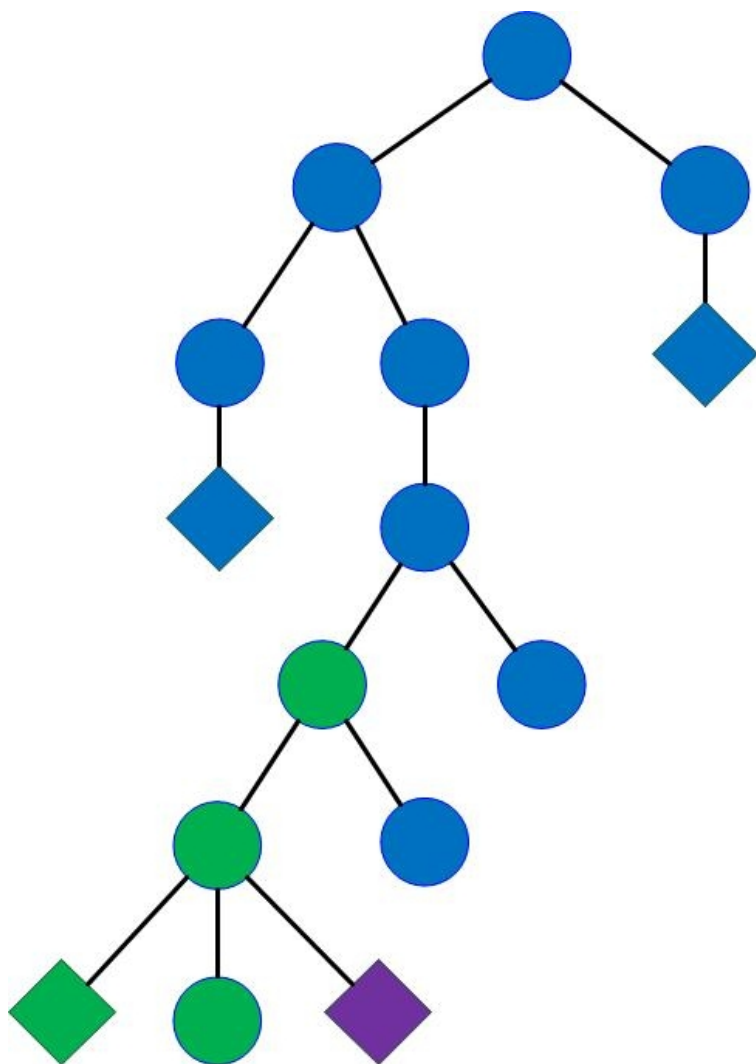
Построение TreePattern



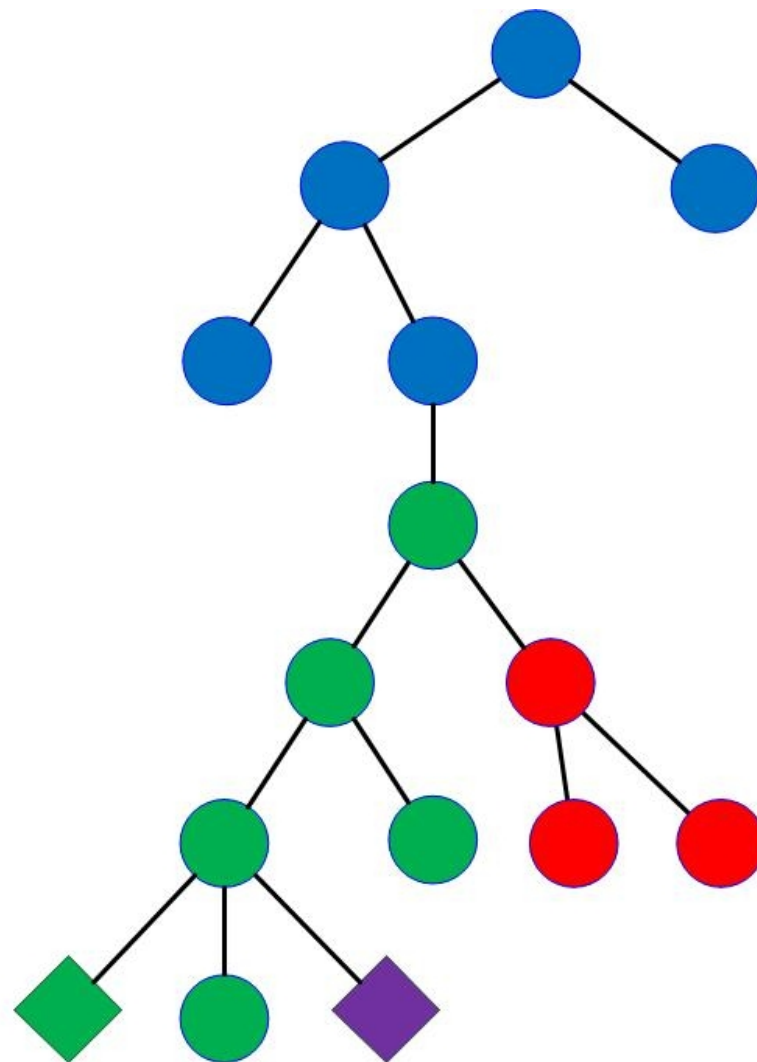
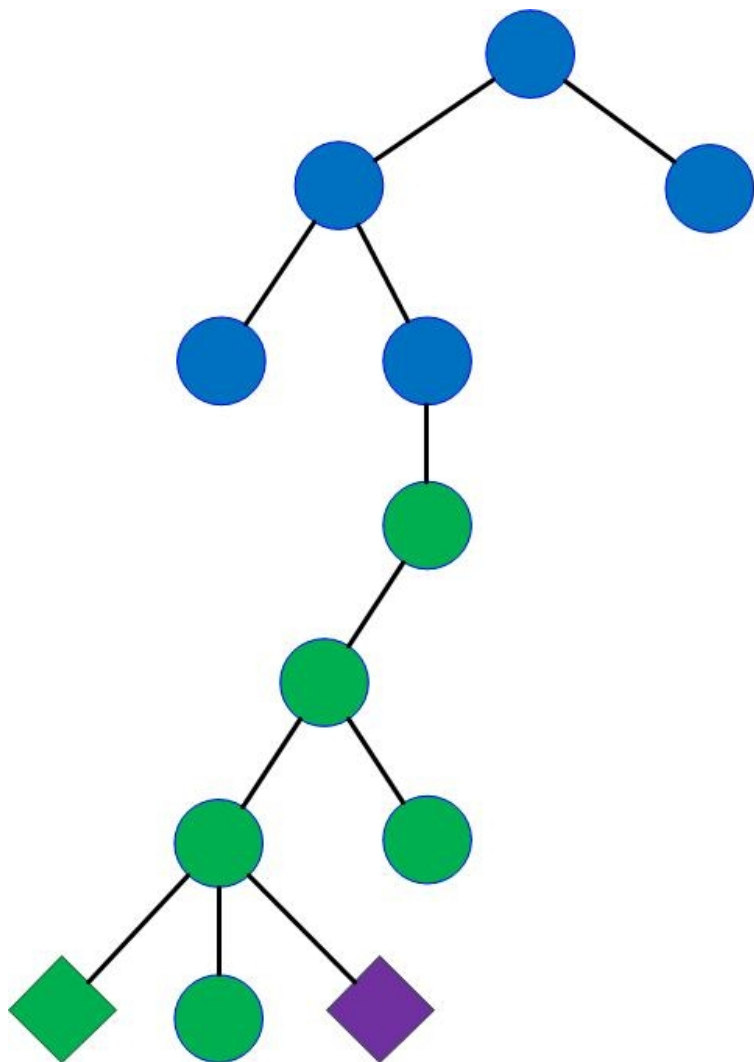
Построение TreePattern



Построение TreePattern



Построение TreePattern



Проблема

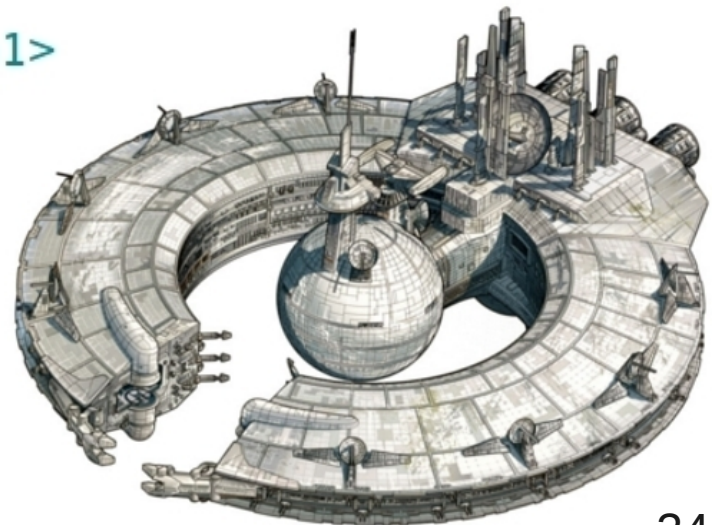
Millennium Falcon



```
<h1>Millennium Falcon</h1>
```

Droid Control Ship

```
<h1><span>Droid Control Ship<span></h1>
```



◆ [Target]

Препятствия

- Сложности с вхождением примеров
- Неправильные страницы

ПОИСК ВХОЖДЕНИЙ

Millennium Falcon



```
<title>StarWars.com | Millennium Falcon</title>
<body>
  <!-- name -->
  <h1>Millennium Falcon</h1>
  <ul>
    <!-- homeworld -->
    <li>Homeworld:<span>Corellia</span></li>
    <!-- size/height -->
    <li>Size:<span>26.7 meters long</span></li>
```

Множественные вхождения

```
<title>StarWars.com | Millennium Falcon</title>
<body>
  <!-- name -->
  <h1>Millennium Falcon</h1>
  <ul>
    <!-- homeworld -->
    <li>Homeworld:<span>Coreellia</span></li>
    <!-- size/height -->
    <li>Size:<span>26.7 meters long</span></li>
```

```
<title>StarWars.com | Leviathan</title>
<body>
  <!-- name -->
  <h1>Leviathan</h1>
  <ul>
    <!-- size/height -->
    <li>Size:<span>600 meters long</span></li>
```



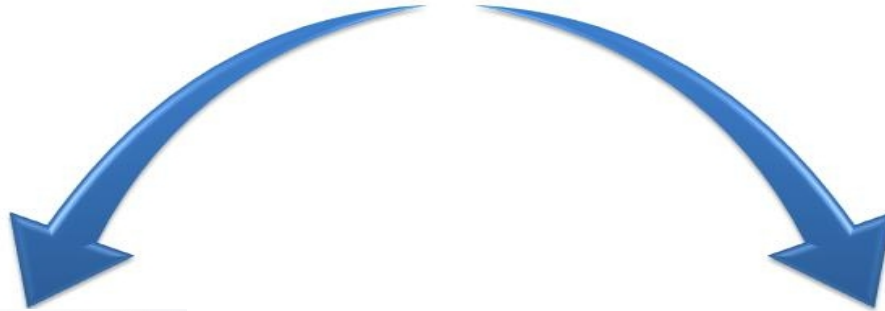
4 варианта

Выбираем лучший

Фильтрация страниц



Все страницы



Применяем
шаблоны



Не
применяем
шаблоны

Глава 4.

ИТОГИ

ИЗМЕНЯЮЩИЙСЯ КОНТЕНТ

Изменяется не везде

Хватает информации для формирования шаблонов

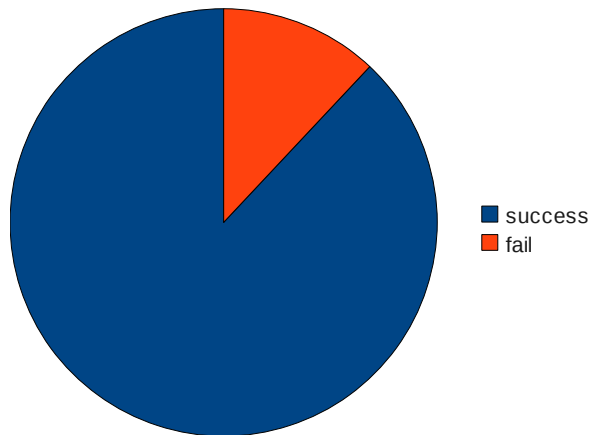
Есть возможность исправить пример

Статистика

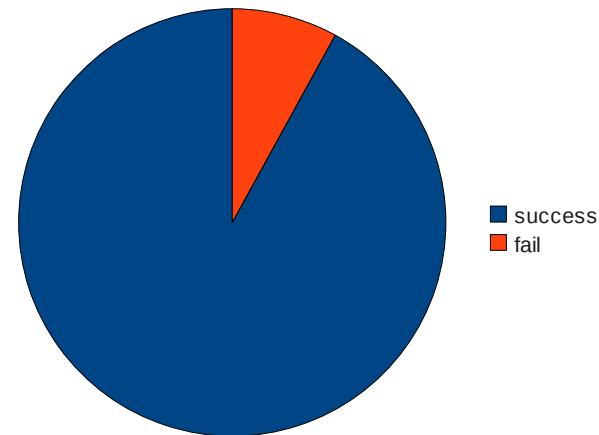
Время создания < 20 мин

Предварительная оценка качества – несколько секунд

Полнота: 88%



Точность: 92%



Нерешенные задачи

- Несколько объектов на странице
- Хорошая устойчивость к неоднородной верстке

Заклучение

Полуавтоматические методы

- По качеству данных и универсальности сравнимы со сбором «в ручную»
- Сложность и время настройки минимальны

P.S. Экономьте Ваше
время!

Вопросы



Нурк Сергей

Разработчик

111033, Россия, Санкт-Петербург,
Свердловская наб., д. 44.

sergeynurk@yandex-team.ru