

Граф знаний

Евгений Чернов

Google's mission



*“Our company mission is to **organize** the world's information and make it universally accessible and useful”*



Понимание текстов



tion and do not even enforce a graph structure. In addition, size is highlighted as an essential characteristic, which is reflected by phrases such as “large networks” or “vast networks” [11], while it remains unclear what “large” means in this context. Färber et al. defined a knowledge graph as an Resource Description Framework (RDF) graph and stated that the term KG was coined by Google to describe any graph-based knowledge base (KB) [7]. Although this definition is the only formal one, it contradicts with more general definitions as it explicitly requires the RDF data model. Pujara et al. did not provide a concise definition, but rather described the characteristics of knowledge graphs. Unlike the other definitions, which focus solely on the inner structure of the KG, they highlighted the importance of an automatic extraction system. In the preface of the 13th International Semantic Web Conference Proceedings (2014), the following statement was published:

Significantly, major companies, such as Google, Yahoo, Microsoft, and Facebook, have created their own “knowledge graphs” that power semantic searches and enable smarter processing and delivery of data: The use of these knowledge graphs is now the norm rather than the exception. [14]

Once again, this highlights the demand for a common definition, because it is necessary to define and differentiate KGs from other concepts in order to make valuable and accurate statements about the introduction and dissemination of knowledge graphs. Furthermore, this ISWC statement proclaims the use of knowledge graphs to be the norm in general, instead of restricting the scope, domain, or application area where KGs can be used beneficially and efficiently. Despite its lack of clarity, this statement seems to have inspired many researchers to submit papers about knowledge graphs in the following conference in 2015².

In the 1980s, researchers from the University of Groningen and the University of Twente in the Netherlands initially introduced the term knowledge graph to formally describe their knowledge-based system that integrates knowledge from different sources for representing natural language [10, 15]. The authors proposed KGs with a limited set of relations and focus on qualitative modeling including human interaction, which clearly contrasts with the idea of KGs that has been widely discussed in recent years.

In 2012, Google introduced the *Knowledge Graph* as a semantic enhancement of Google’s search function that does not match strings, but enables searching for “things”, in other words, real-world objects [18]. Although the blog post does not provide any implementation details, it has been cited more than 100 times according to Google Scholar³. Since 2012, the term knowledge graph is also used to describe a family of applications. Frequently mentioned implementations are DBpedia, YAGO (Yet Another Great Ontology), Freebase, Wikidata, Yahoo’s semantic search assistant tool Spark, Google’s Knowledge Vault, Microsoft’s Satori and Facebook’s entity graph [7, 14, 16, 11]. Those applications differ in their characteristics, such as architecture, operational purpose, and technology used, which makes it difficult to find a consensus and to create a definition of knowledge graph. The lowest common denominator of the listed open source applications is their use of Linked Data, whereas hardly any proven information is available about Satori and the entity graph.

In addition, the more specific term *enterprise knowledge graph* is used by a few smaller companies, for example, SindiceTech⁴ and the Semantic Web Company [3]. Both companies seek to describe a similar model that extracts and stores diverse enterprise data in a triple store and analyzes it by using machine learning techniques in order to acquire new knowledge from the data and to reuse it in other applications.

Набор слов и предложений

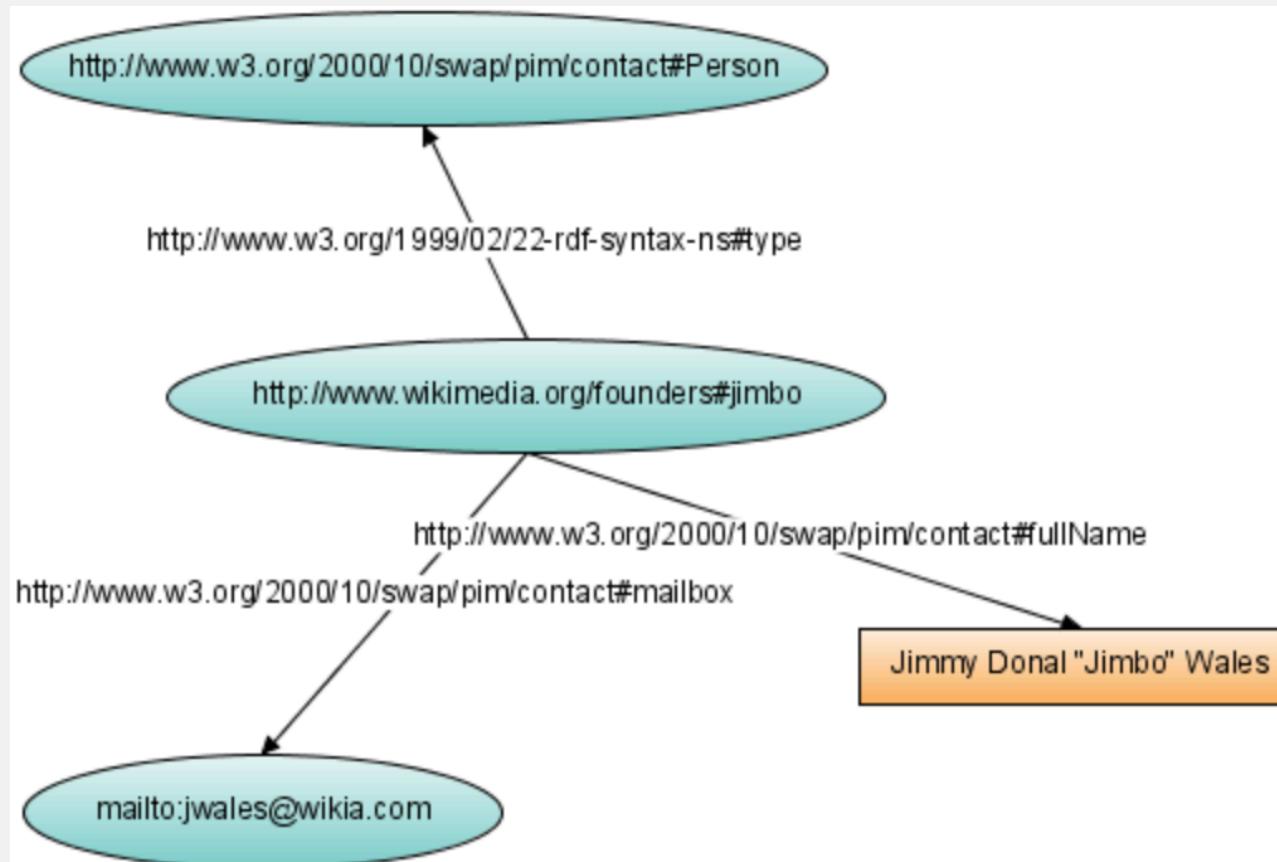


Как понять смысл?



- Semantic Web
- *Tim Berners-Lee, 1998*
- *WWW -> GGG (Giant Global Graph)*

Semantic Web



- Индентификатор объектов – URI
- Онтология

Онтология



- Экземпляры
 - *Аватар*
- Понятия
 - *Фильм*
- Атрибуты
 - *Название*
 - *Рейтинг*
 - *Сборы по миру*
- Отношения
 - *Аватар* -> *director* -> *Джеймс Кэмерон*

Микроформат



Добавляем смысл в HTML

```
<div>
  <div>Василий Пупкин</div>
  <div>Рога и Копыта</div>
  <div>495-564-1234</div>
  <a href="http://example.com/">Мой сайт</a>
</div>
```

```
<div class="vcard">
  <div class="fn">Василий Пупкин</div>
  <div class="org">Рога и Копыта</div>
  <div class="tel">
    <span class="type">Work</span>
    <span class="value">495-564-1234</span>
  </div>
  <a class="url" href="http://example.com/">Мой сайт</a>
</div>
```

Микроформаты



- **hCard** – организации и люди
- **hCalendar** – события
- **adr** – почтовые адреса
- **geo** – географические координаты
- **hProduct** – товары
- **hRecipe** – кулинарные рецепты приготовления блюд

Микроразметка



- schema.org – более общий и простой способ аннотирования текстов
- Поддержка с HTML5
- Появился в 2011 году
- Разработан Google, Yahoo!, Bing

```
<div itemscope itemtype="http://schema.org/ScholarlyArticle">
    <h1 itemprop="name">Заголовок статьи</h1>
    <div itemprop="author">ФИО автора</div>
    <div itemprop="articleBody">
        Текст статьи
    </div>
</div>
```

Микроразметка: Кинопоиск



The movie poster for "Avatar" features Jake Sully's face on the left and the Na'vi Jake on the right, with floating mountains and a bioluminescent environment in the background.

Трейлер

+ Буду смотреть

все папки (159415)

Аватар IMAX

Avatar

▶ от 9 ₽ от 99 ₽

Скидка 90% на первые три покупки для подписчиков

год 2009

страна США

слоган «Это новый мир»

режиссер Джеймс Кэмерон

сценарий Джеймс Кэмерон

продюсер Джеймс Кэмерон, Джон Ландау, Брук Бретон, ...

оператор Мауро Фиоре

композитор Джеймс Хорнер

художник Рик Картер, Роберт Стромберг, Мартин Лэйн, ...

монтаж Джеймс Кэмерон, Джон Рефуа, Стивен Е. Ривкин

жанр фантастика, боевик, драма, приключения, ... слова

бюджет \$237 000 000

сборы в США \$749 766 139

сборы в мире + \$1 994 570 654 = \$2 744 336 793

сборы в России \$119 903 638

DVD в США \$190 196 351

зрители 97.3 млн, 27.6 млн, 16.5 млн, ...

премьера (мир) 10 декабря 2009, ...

премьера (РФ) 17 декабря 2009, «Двадцатый Век Фокс СНГ»

ре-релиз (РФ) 26 августа 2010

В главных ролях:

Сэм Уортингтон
Зои Салдана
Сигурни Уивер
Стiven Энг
Мишель Родригес
Джованни Рибири
Джоэль Мур
Си Си Эйч Паундер
Уэс Стыди
Лас Алонсо
...

Роли дублировали:

Александр Ноткин
Мария Цветкова-Овсянникова
Елена Шульман
Мария Кузнецова
Валерий Соловьев
...

показать всех »

Извлечение структурированной информации



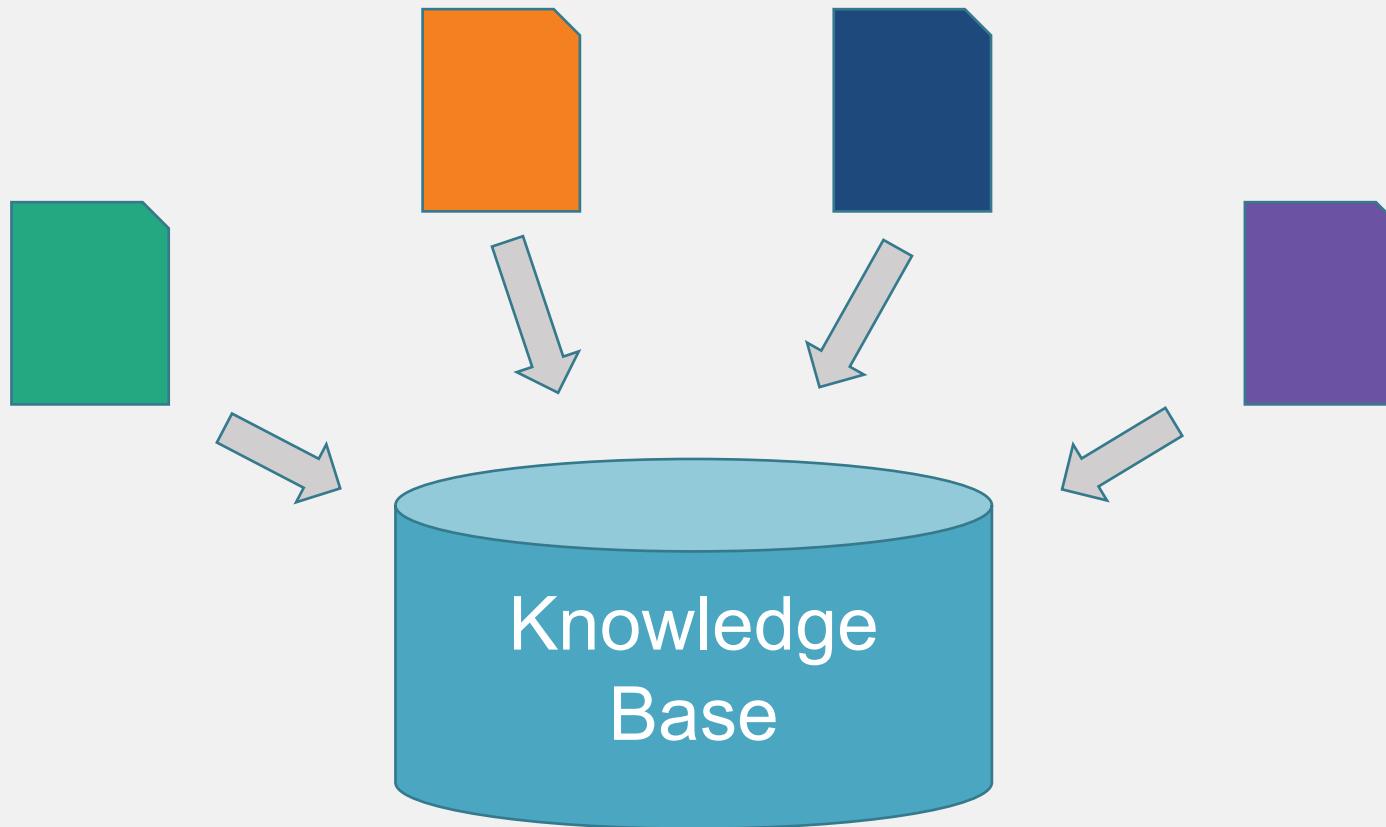
```
d class="type">слоган</td><td style="color: #555">&laquo;Это новый мир&raquo;</td></tr>
d class="type">режиссер</td><td itemprop="director" > Джеймс Кэмерон</td></tr>
<a href="/lists/navigator/sci-fi/?quick_filters=films">фэнтези</a>, <a href="
<tr class="en">
<td class="type">бюджет</td>
<td class="dollar"><div style="position: relative">
<a href="/film/251733/box/" title="">$237 000 000</a>
</div></td>
tr>
<tr>
```

description	
Землян на планете Пандора, где корпорации добывают редкий минерал, имеющий огромное значение для выхода Земли из энергетического кризиса.	
director	
@type	Person
name	Джеймс Кэмерон
producer	
@type	Thing
name	Джеймс Кэмерон, Джон Ландау, Брук Бретон, ...
musicBy	
@type	Thing
name	Джеймс Хорнер
actors	
@type	Person
name	Сэм Уортингтон
actors	

База знаний



Цель: единая база всех объектов и их связей



Дополнительная информация об объекте



Поиск репин

Найти

Интернет Соцсети beta Картинки Видео Новости Ответы

Репин, Илья Ефимович — Википедия
ru.wikipedia.org/wiki/Репин,_...
Илья Ефимович Репин — русский живописец, педагог, профессор, действительный член Императорской Академии художеств. Уже с самого начала своего творческого пути, с 1870-х годов, Репин стал одной из ключевых фигур русского реализма. Художнику удалось решить задачу отражения в живописном...

Происхождение. Детство... Пенсионерская поездка за...
Первый петербургский период... Московский период (1877—1882)
Первая семья Второй петербургский период...

Картинки по запросу «репин» 16 тыс. результатов

К Репин Илья Ефимович — биография художника, личная жизнь, картины culture.ru/persons/8244/ilya-repin
Полная биография художника Ильи Репина. Личная жизнь, фотографии, интересные факты из жизни, картины художника на портале «Культура.РФ»

M Репин: картины, биография. Произведения Ильи Ефимовича Репина muzei-mira.com/biografia_...
Илья Ефимович Репин. Его биография и подробный список картин с названиями. Описание картин художника.

24 Илья Репин — биография, фото, личная жизнь, картины, произведения ... 24smi.org/celebrity/6329-ilia-...
Биография Ильи Репина: личная жизнь, картины, произведения, творчество, «Иван Грозный и сын его Иван 16 ноября 1581 года», «Бурлаки на Волге», «Не ждали», «Запорожцы», «Садко», «Крестный ход в...

★★★★★ 8 из 10 | 17 голосов

Б ИЛЬЯ РЕПИН. Биография и картины художника Ильи Репина (Ильи Ефимовича Репина) ru.wikipedia.org/...
Илья Ефимович Репин — русский живописец, педагог, профессор, действительный член Императорской Академии художеств. Уже с самого начала своего творческого пути, с 1870-х годов, Репин стал одной из ключевых фигур русского реализма. Художнику удалось решить задачу отражения в живописном...

Илья Репин
Русский живописец

Илья Ефимович Репин — русский живописец, педагог, профессор, действительный член Императорской Академии художеств. Википедия
Родился: 5 августа 1844 г., Чугуев, Украина
Умер: 29 сентября 1930 г. (86 лет), Репино, СССР
Родители: Репина Татьяна Степановна
Дети: Юрий Репин
В браке с: Репина Вера Алексеевна

Работы

Ещё

Запорожцы Бурлаки на Волге Иван Грозный и сын его Иван 16 ноября 1581 года Славянские композиторы Крестный ход в Курской...

Смотрите также

Ещё

Ответы на вопросы



поиск в каких фильмах играл высотский Найти

Интернет Соцсети beta Картинки Видео Новости Ответы

Владимир Высоцкий — Фильмы и сериалы

Место встречи изменит... 1979	Маленькие трагедии 1980	Служили два товарища 1968	Короткие встречи 1968	Плохой хороший человек 1973	Увольнение на берег 1962	Сказ про то, как царь ... 1976	Карьера Димы Горина 1961	Наш дом 1965

поиск что посмотреть в праге Найти

Интернет Соцсети beta Картинки Видео Новости Ответы

Прага — Достопримечательности

Карлов мост	Собор Святого Вита	Вацлавская площадь	Пражский Град	Староместская площадь	Пороховая башня	Королевский сад

Ответы на вопросы



@ ПОИСК когда родился пушкин x

Интернет Соцсети beta Картинки Видео Новости Ответы

Александр Пушкин > Дата рождения

6 июня 1799 г.



@ ПОИСК высота килиманджаро x

Интернет Соцсети beta Картинки Видео Новости Ответы

Килиманджаро > Высота

5895 м



Exploratory search



Фильмы

Человек дождя 1988	Последний самурай 2003	Интервью с вампиром 1994	Грань будущего 2014	Несколько хороших пар... 1992
---------------------------------------	---	---	--	--

Ещё

Смотрите также

Николь Кидман	Пенелопа Крус	Ребекка Де Морнэй	Брэд Питт	Джон Траволта
-------------------------------	-------------------------------	-----------------------------------	---------------------------	-------------------------------

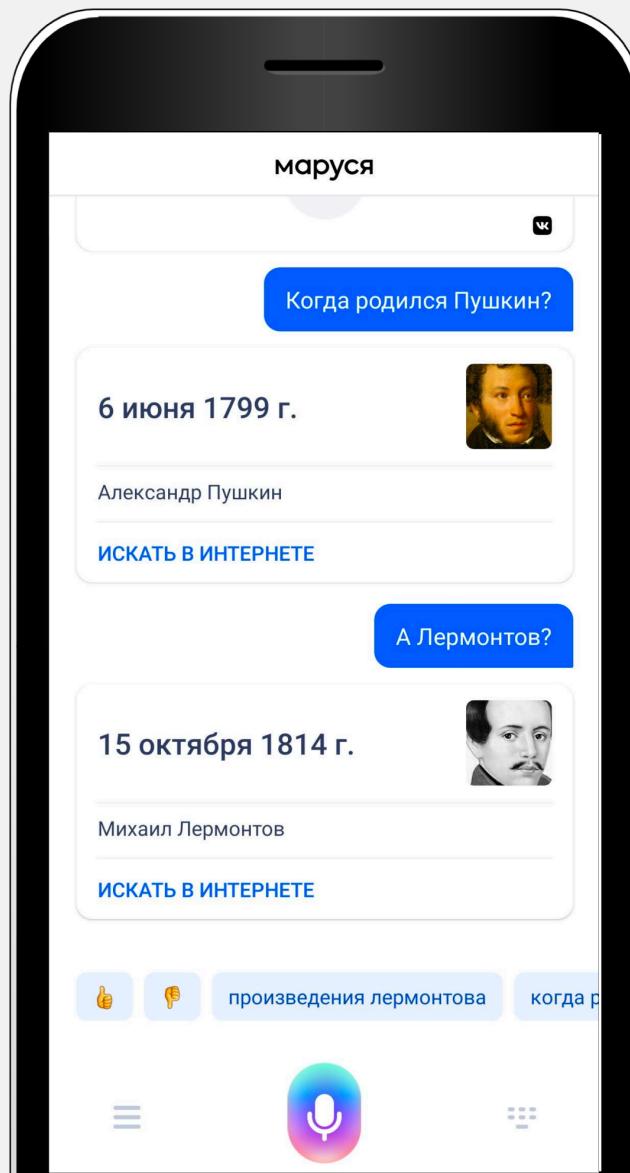
Ещё

Рекомендации и фильмы в карточке Тома Круза

Голосовые помощники



Добавляется
контекст



Готовые базы знаний





Проект по структурированию данных из Википедии

FreeBase

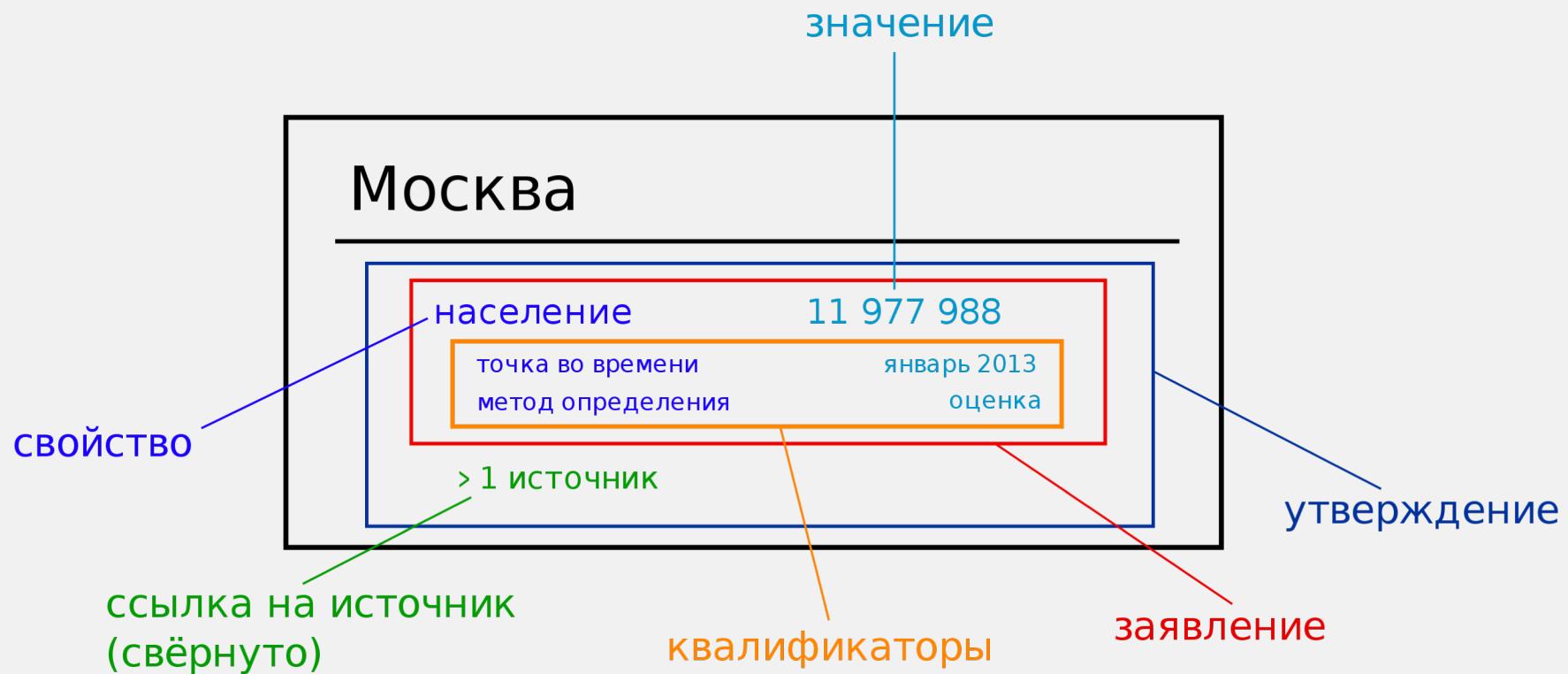


- Сообщество собирало знания из различных источников
- Google купил в 2015 году

Wikidata



- Совместно редактируемая база знаний
- Запущена 2012 году Фондом Викимедия



Wikidata: Пушкин



Отношение

sex or gender

male

▶ 3 references

edit

+ add value

country of citizenship

Russian Empire

▶ 1 reference

edit

+ add value

Значения

Приоритеты

name in native language

Александр Сергеевич Пушкин (Russian)

▼ 0 references

edit

+ add reference

Александръ Сергѣевичъ Пушкинъ (Russian)

writing system

Russian orthography before 1918

▼ 0 references

edit

+ add reference

+ add value

Статистика по открытым базам знаний



Источник	DBpedia	Freebase	Wikidata
Частота обновления	раз в год	не обновляется	еженедельно
Объем ¹	8.1 GB	31 GB	9.5 GB
Источники данных	википедия (автомат.)	википедия, другие источники, команда freebase	пользователи и боты
Контроль качества	зависит от википедии и парсеров	команда freebase	сообщество

Статистика по открытым базам знаний



Источник	DBpedia	Freebase	Wikidata
Идентификатор объекта	Karlsruhe	/m/0qb1z	Q1040
Описание объекта	длинное	длинное	короткое
Источники фактов	не указаны	не указаны	указаны
Релевантность объектов	нет	есть	нет
Релевантность отношений	нет	нет	есть

Тематические базы знаний



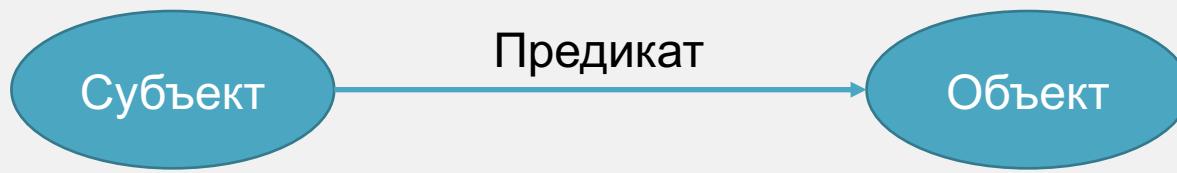
OPEN LIBRARY

 MusicBrainz



КиноПоиск

Resource Description Framework



- **RDF** - семейство спецификаций для представления данных
- **N-Triples** (триплеты) — текстовый формат, используемый для хранения и передачи графов RDF

<subject> <predicate> <object>

RDF: пример



```
<http://mythology.Greek.org/#Cronus>
```

```
<http://www.example.org/schemas/rel/fatherOf>
```

```
<http://mythology.Greek.org/#Zeus>
```



SPARQL Protocol and RDF Query Language

PREFIX

префиксные объявления - служат для указания сокращений универсальных идентификаторов ресурса (URI), используемых в запросе.

FROM ...

источники запроса - определяют какие RDF-графы запрашиваются.

SELECT ...

состав результата - определяет возвращаемые элементы данных.

WHERE {...}

шаблон запроса - определяет, что запрашивать из набора данных.

ORDER BY ...

модификаторы запроса - ограничивают, упорядочивают,

преобразуют результаты запроса

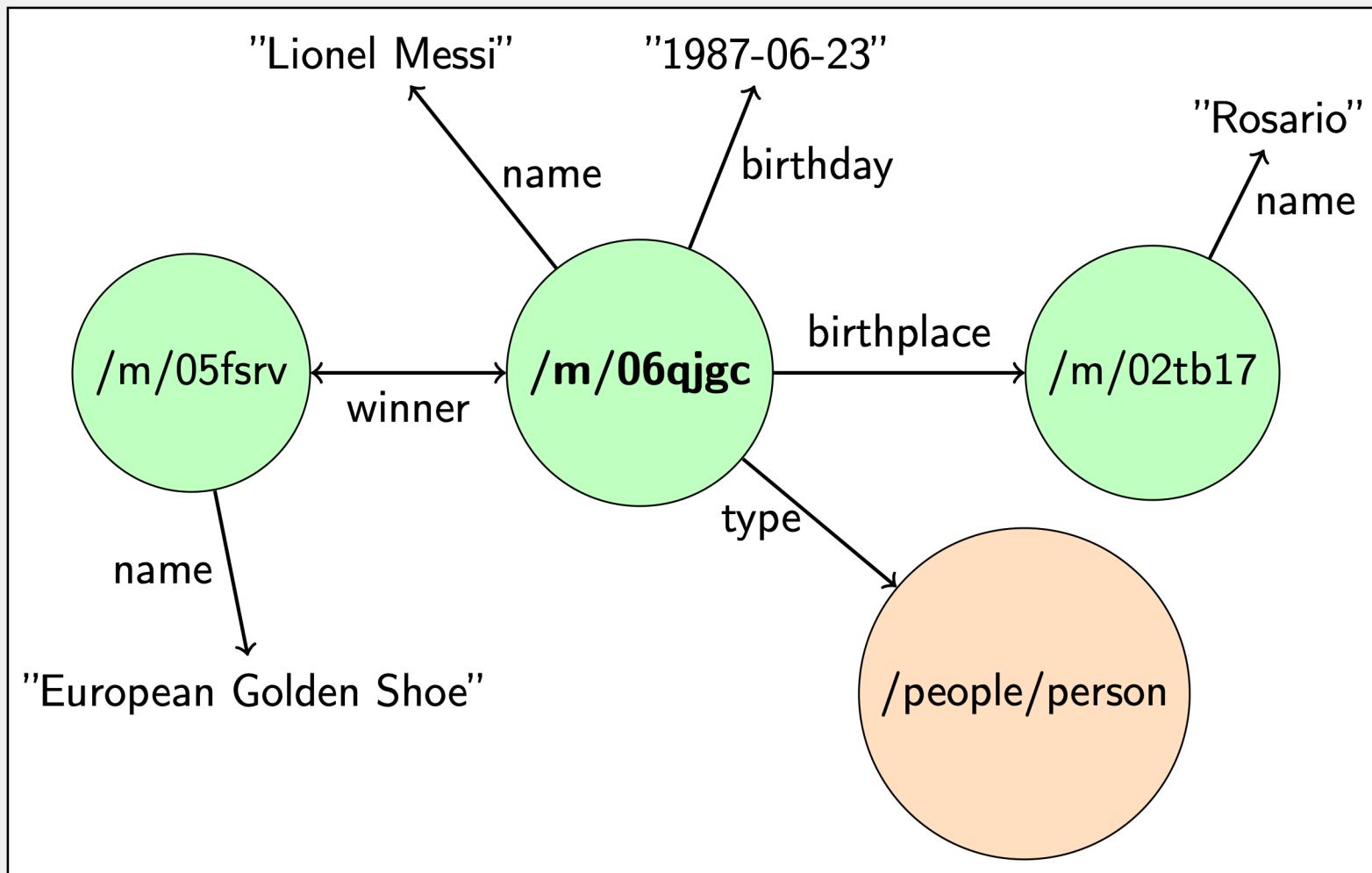
SPARQL



```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?name ?email
WHERE {
    ?person a foaf:Person.
    ?person foaf:name ?name.
    ?person foaf:mbox ?email.
}
```

Запрос на имена и email всех людей в базе

Граф знаний



Построение собственной базы знаний

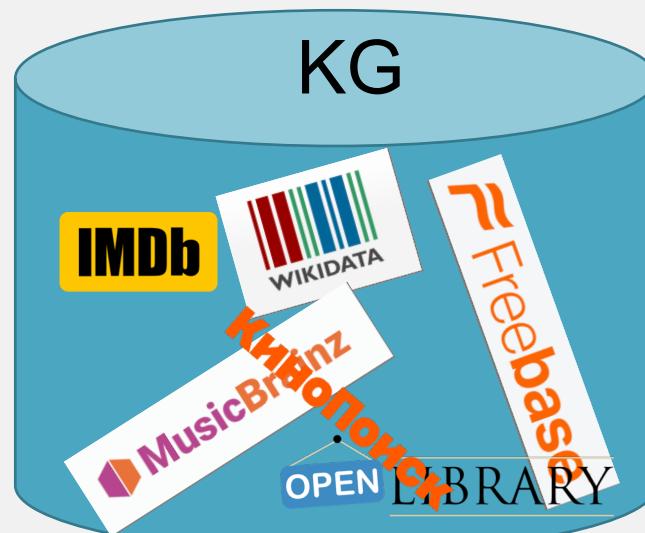


Проблема:

- Открытые базы данных неполны

Цель:

- Слияние различных баз знаний в единую базу
 - *Named Entity Linking (NEL)*



Wikidata + КиноПоиск



Преобразуем данные к единому формату (N-Triples)

- Викиданные \approx 720 GB
- КиноПоиск \approx 230 GB

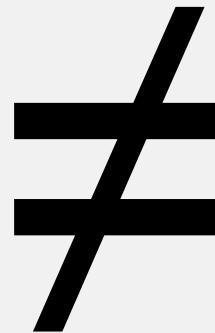
Склейка дублей



КиноПоиск

Один объект реального мира должен быть одним объектом в графе знаний

Склейка дублей



Александр Козлов
Alexander Kozlov

К сожалению,
ФОТО
отсутствует

■ дата рождения	-
■ место рождения	-
■ жанры	драма, комедия
■ всего фильмов	1, 2013

+ Избранное



КиноПоиск

Объединение баз: проблемы



- Существует как минимум два человека с именем Брэд Питт (актёр и боксёр).
- Более двухсот фильмов и сериалов под названием «Мост».
- Более трёх тысяч Иванов Ивановых связанных с кино

Объединение баз: решение



- Генератор кандидатов для связывания.
- Классификатор. Решает нужно ли склеивать между собой пару объектов.
- Дополнительная склейка. Доклеиваем то, что не получилось на предыдущем шаге

Объединение баз: генератор



Основная идея

Имена объектов, которые могут быть связаны должны быть близки

- Полные совпадения подстрок, но с возможностью перестановки слов
- Русскоязычные и иностранные имена
- Алиасы объектов (перенаправления Википедии)

Данные для обучения



Все необходимое уже есть в открытых графах знаний

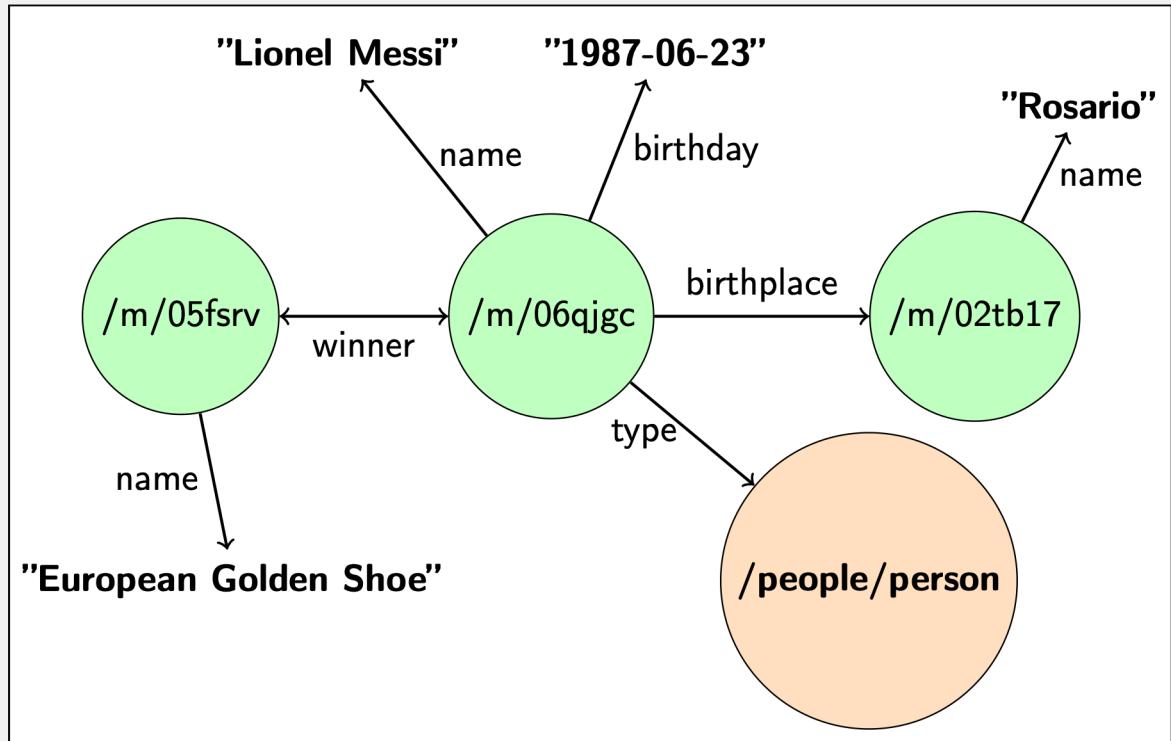
Wikidata:

- Ссылки на фильмы и сериалы КиноПоиска — **180 228** объектов
- Ссылки на персон КиноПоиска — **98 668** объектов

Признаки



Однаковые
объекты имеют
похожий контекст



Контекст объекта — инцидентные ему строки, имена соседей и типы



Вычисление признаков



Совпадения строк

- Полные совпадения
- Совпадения слов

Группировка

- Объединяются в группы по отношениям
- Сваливаются в одну кучу

Считаем абсолютные и относительные совпадения по группам

Примеры признаков



Лионель Месси

Name: "Lionel Messi"

Birthday: "June 24, 1987"

Football club: "FC Barcelona"

Лайонел Ричи

Name: "Lionel Richie"

Birthday: "June 20, 1949"

- Полное совпадение имени = 0,
- Количество совпавших слов в имени = 1 (Lionel),
- Коэффициент Жаккара для имени = 0.333,
- ...
- Количество совпавших слов в FC = \emptyset
- Всего полных совпадений = 0,
- Всего совпадений слов = 2 (Lionel, June).

Лучшие признаки



Лучшие признаки



XGBoost

250 деревьев
высота 4

Метрика	Значение
Precision	0.961531
Recall	0.963220
F_1	0.962375
AUC	0.999589

«Большие» объекты



Города, страны, жанры, рода занятий — не склеиваются

- Страдает согласованность данных
- Похожих объектов нет в обучающем множестве

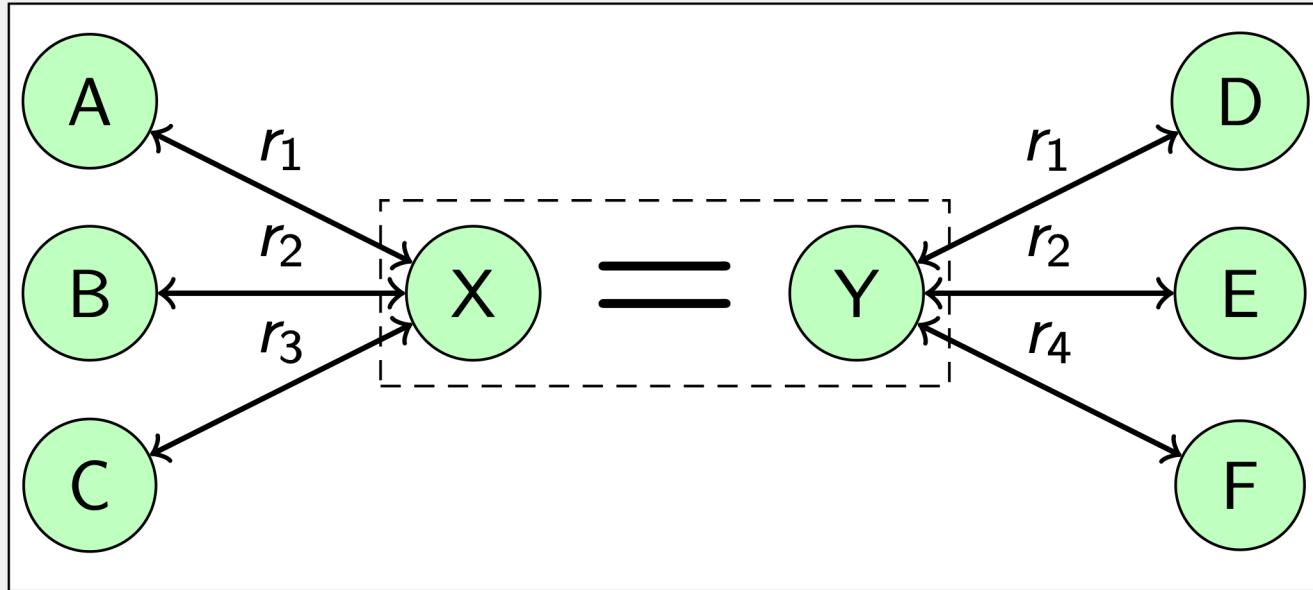
ХавьёрÁнхель Энсйнас Бардём — испанский актёр. Обладатель премии «Оскар» за лучшую мужскую роль второго плана, а также «Золотого глобуса» за фильм «Стариким тут не место» и приза Каннского фестиваля за лучшую мужскую роль за роль в мексикано-испанском фильме «Бьютифул». [Википедия](#)

Родился: 1 марта 1969 г. (50 лет), [Лас-Пальмас-де-Гран-Канария, Испания](#), [Лас-Пальмас-де-Гран-Канария, Испания](#), [Лас-Пальмас-де-Гран-Канария, Испания](#)

Рост: 181 см

Города «Лас-Пальмас-де-Гран-Канария» и страны «Испания» из разных источников не склеились

Likelihood Ratios



$$H_1 : P(D | A) = p = P(D | \neg A)$$

$$H_2 : P(D | A) = p_1 \neq p_2 = P(D | \neg A)$$

Информационная карточка



W **Месси, Лионель — Википедия**
ru.wikipedia.org/wiki/Месси,_Лионель

Лионэль Андре́с Месси — аргентинский футболист, нападающий футбольного клуба «Барселона», с 2011 года — капитан национальной сборной Аргентины. Лучший бомбардир в истории «Барселоны» и сборной Аргентины. Считается одним из лучших футболистов современности и одним из лучших игроков всех времён.

[Биография](#) [Личная жизнь](#)
[Карьера](#) [Статистика выступлений](#)
[Сборная Аргентины](#) [Достижения](#)

Видео по запросу «лео месси» 32 тыс. результатов



КАК ЛЕО МЕССИ
ЗАЩИЩАЕТ СВОИХ
Youtube

Лео Месси и Луис
Суарес - Лучшие
Youtube

МЕССИ О РОНАЛДУ ► Досье
НЕОЖИДАННОЕ
Youtube

Vk.com
25 авгу

Лео Месси
Аргентинский футболист



Лионэль Андре́с Месси — капитан испанского клуба национальной сборной Аргентины «Барселоны» и сборной Аргентины. Считается одним из лучших футболистов современности и одним из лучших игроков всех времён.

[Википедия](#)

Лионель Месси

Источники:

https://ru.wikipedia.org/wiki/Месси,_Лионель

https://en.wikipedia.org/wiki/Lionel_Messi

<https://www.kinopoisk.ru/name/2317372>

Модель показов карточки



Признаки классификатора показов

- Тестовые (как много совпадений в тексте документа)
- Поведенческие (как часто пользователи кликают в документ-объект по запросу)
- Позиция документа в выдаче (намного дальше, чем первая страница)

Хранилище: требования



- Обработка запросов вида: $(s, *, *)$, $(s, p, *)$, (s, p, o)
- Возможность изменения схемы хранения данных графа
- Быстрое построение с нуля
- Время ответа сервиса менее 100 мс на 99% запросов

Хранилища: готовые решения



- Реляционная база данных (таблицы для узлов и ребер)
- Redis (in-memory)
- Apache Jena (долгая индексация)
- Cayley (not production ready)

Хранилище: ключевое наблюдение



Конкретные значения строк нужны только в самом конце

/m/0283x author /m/034hwx	\rightarrow	$h_1 h_2 h_3$
/m/0283x birthplace /m/0dy19		$h_1 h_4 h_5$

- ▶ $h_1 = \text{hash}(/m/0283x),$
- ▶ $h_2 = \text{hash}(\text{author}),$
- ▶ $h_3 = \text{hash}(/m/034hwx),$
- ▶ $h_4 = \text{hash}(\text{birthplace}),$
- ▶ $h_5 = \text{hash}(/m/0dy19).$

Хранилище: разделение данных



Структура графа

Упорядоченный файл с хешированными строками троек.

- ▶ $h_1 h_2 h_3,$
- ▶ \dots
- ▶ $h_n h_m h_k.$

Строки

Однородный массив строк.

$|s_1|, s_1, |s_2|, s_2, \dots, |s_n|, s_n.$

Индекс строк

Обратное преобразование хеш → номер строки.

Схема хранения



•

- Структура графа ≈ 35 GB (in-memory)
- Индекс строк ≈ 7 GB (in-memory)
- Строки ≈ 30 GB (in-memory + SSD)

Хранилище: поиск



Запросы $(s, *, *)$ и $(s, p, *)$

Два бинарных поиска на запрос.

- ▶ Начало блока: $(s, 0, 0)$ или $(s, p, 0)$.
- ▶ Конец блока: (s, ∞, ∞) или (s, p, ∞) .

Запрос (s, p, o)

Один бинарный поиск на запрос.

Отмечайтесь и оставляйте отзыв

Спасибо за
внимание!

Евгений Чернов

e.chernov@corp.mail.ru