

Поисковые подсказки

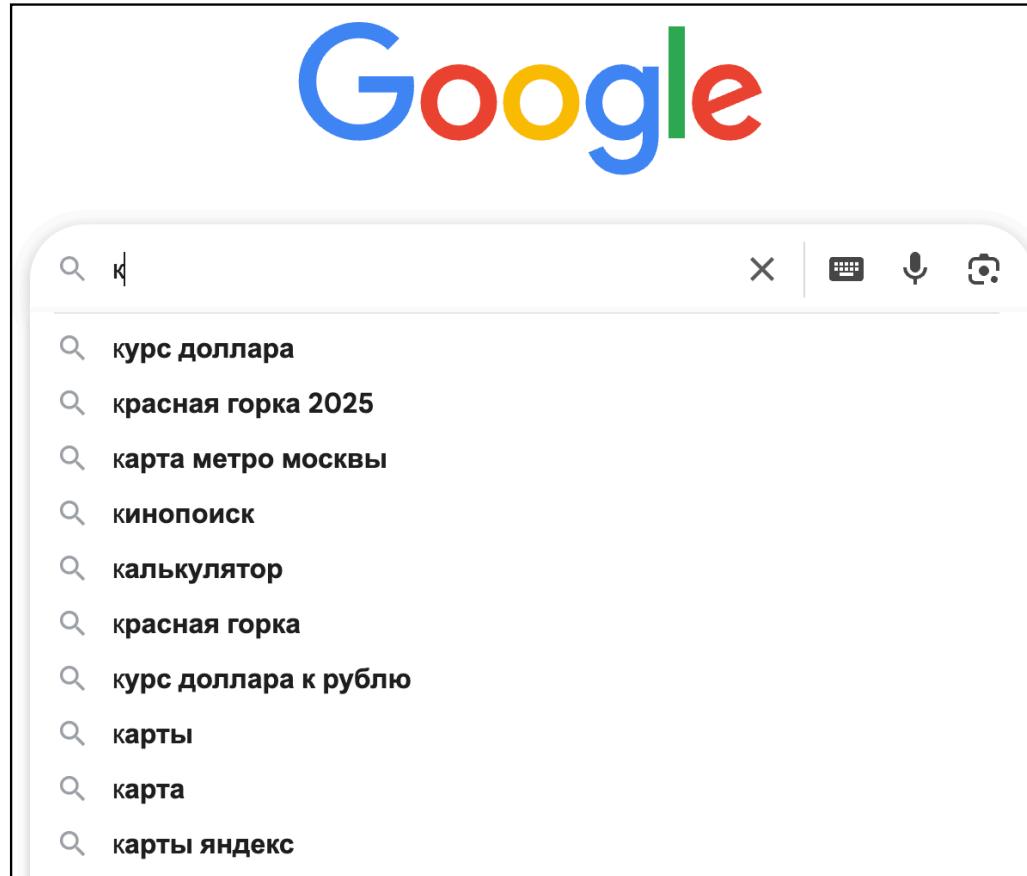
Андронов Дмитрий, 19.05.2025, AI Masters

План лекции

- Постановка задачи поисковых подсказок
- Классическая система поисковых подсказок
- Генеративная постановка задачи

Другие задачи поисковых систем

Поисковые подсказки



Другие задачи поисковых систем

Исправление опечаток

Search bar: кроссовки мушские

Search filters: Все, Картинки, Покупки, Короткие видео, Новости, Видео, Веб-версия, Ещё, Инструменты

Category suggestions: New Balance, Фирменные, Найк, Кожаные, Адидас, Зимние, Летние, Крутые, 2025

Text below results: Показаны результаты по запросу **кроссовки мужские**
Искать вместо этого **кроссовки мушские**

Другие задачи поисковых систем

Синонимы

стиралка

Все Картинки Покупки Видео Короткие видео Новости Карты : Ещё Инструменты

Это Купить С сушкой Самсунг Размеры Баш Индезит Мини Отжимает

Результаты: Московский, Поселение, Москва Точное местоположение

M.Видео <https://www.mvideo.ru> > ... > Стиральные машины

Стиральные машины автомат купить в ...

Стиральные машины по низкой цене в интернет-магазине М.Видео. Заказать стиралку автомат с доставкой по телефону 8 (800) 600-777-5.

От 3 100,00 ₽ до 719 999,00 ₽ · 4,7 ★★★★★ (28 422)

OZON <https://www.ozon.ru> > ... > Крупная бытовая техника

купить стиральную машину на OZON по низкой цене

Стиральные машины – покупайте на OZON по выгодным ценам, быстрая и бесплатная доставка, оригинальные товары, гарантия, бонусы, рассрочка и кэшбэк, ...

От 76,00 ₽ · 4,7 ★★★★★ (26 109)

DNS <https://www.dns-shop.ru> > catalog > стиральные-машин ...

Стиральные машины купить в интернет-магазине ...

Купить Стиральные машины по самым выгодным ценам в интернет-магазине DNS. Широкий выбор товаров и акций. В каталоге можно ознакомиться с ценами, отзывами, ...

От 7 299,00 ₽ до 572 999,00 ₽ · 4,5 ★★★★★ (186 420)

Другие задачи поисковых систем

Блендинг виджетов

Яндекс куртка для собак

поиск

нейро картинки видео карты товары переводчик все

Товары Цена Бренд

Купить: куртка для собак

728 ₽ Попона на молнии для собак VitaVet market.yandex.ru	3 550 ₽ 8 399 ₽ Комбинезон дождевик для собак ASMPET... market.yandex.ru	1 510 ₽ Комбинезон-дождевик для собак Gamma... market.yandex.ru	2 250 ₽ Комбинезон-плащ для собак крупных пород... market.yandex.ru
---	--	---	---

Новая летняя коллекция уже на сайте Street Beat!

street-beat.ru Новая летняя коллекция уже на сайте Street Beat! Промо 🎉
Летние новинки от известных брендов уже на сайте. Выбирайте свой стиль!
★ 4,3 · 1,5K отзывов на магазин
Контактная информация · +7 (495) 105-XX-XX Показать ·
пн-чт 10:00-22:00, пт,сб 10:00-23:00, вс 10:00-22:00 · м. ЦСКА · Москва

Межсезонная распродажа
Тысячи товаров со скидками в Street Beat

Оплата Долями
Оплачивайте покупки частями без комиссий и переплат

Для мужчин
Новая мужская коллекция уже на сайте Street Beat.

Для женщин
Новая женская коллекция уже на сайте Street Beat.

Подарочный сертификат
Подарите электронный сертификат для покупок в Street Beat

Кеды
Большой выбор кед на сайте Street Beat

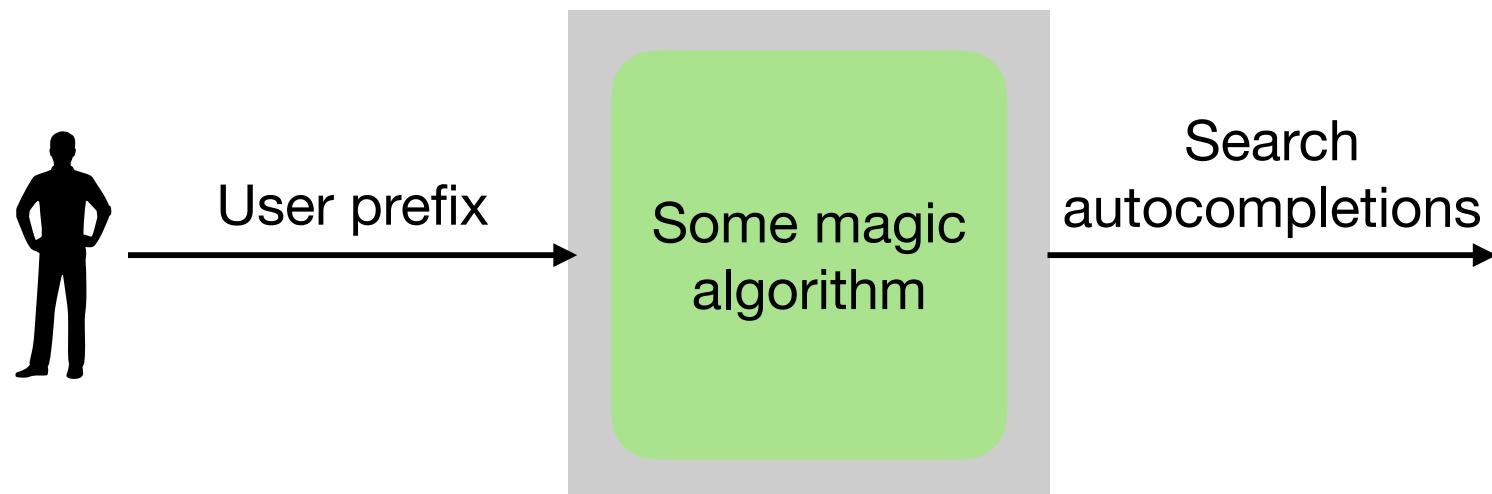
Постановка задачи поисковых подсказок

Постановка задачи

На примере маркетплейса

- Есть работающая поисковая система
- Пользователи используют нашу систему: вводят поисковые запросы, кликают, добавляют в корзину и покупают товары
- Необходимо построить систему поисковых подсказок:
 - Пользователь вводит префикс
 - Система предлагает набор подсказок

Система поисковых подсказок



Примеры систем поисковых подсказок

Везде X 🔍

кронштейн капсулы кофе колбаса кондиционер

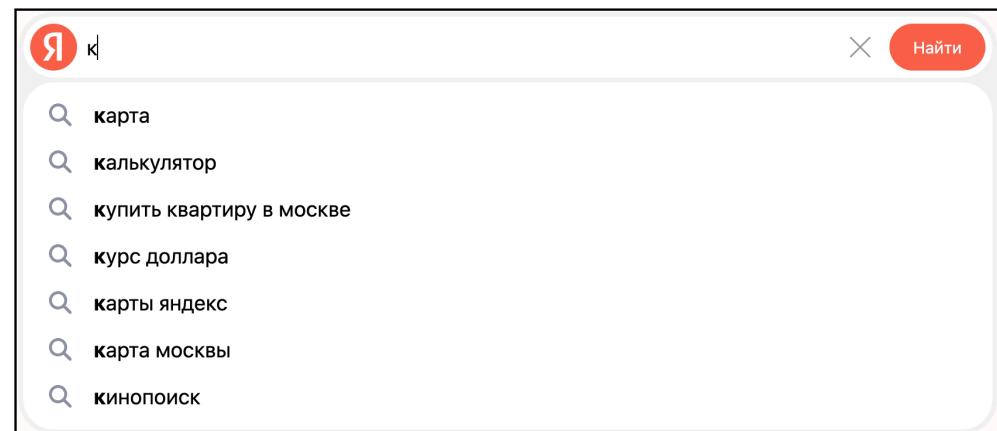
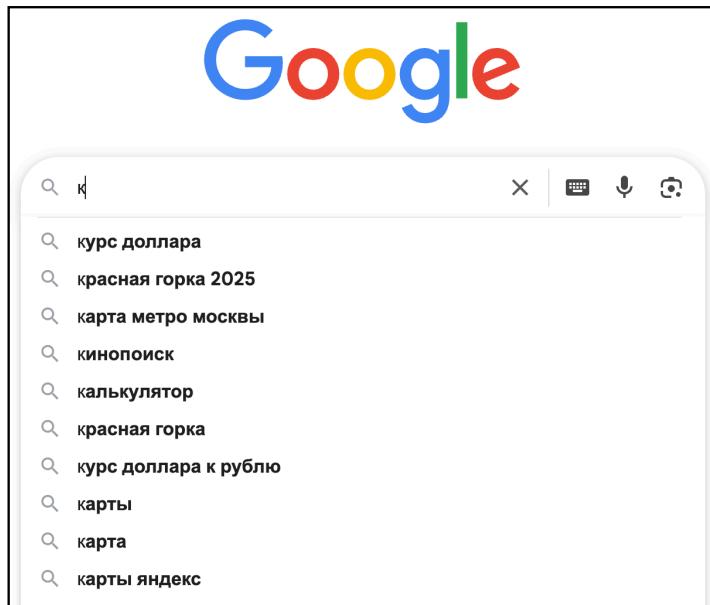
cola колбаса Быстрая доставка FRESH >

- 🔍 кронштейн для телевизора настенный
- 🔍 капсулы для посудомоечной машины
- 🔍 капсулы для стирки
- 🔍 кофе в зернах 1 кг
- 🔍 колбаса
- coffee Кофе в зернах Продукты питания
- kolbas Kolбасы Продукты питания
- Jardin Jardin Бренд
- Klinskiy Клинский Бренд

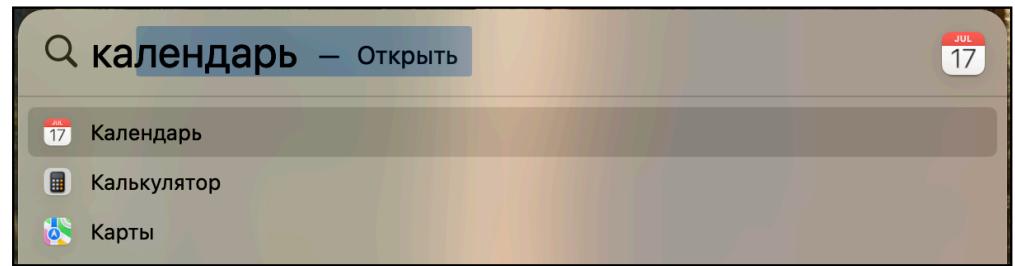
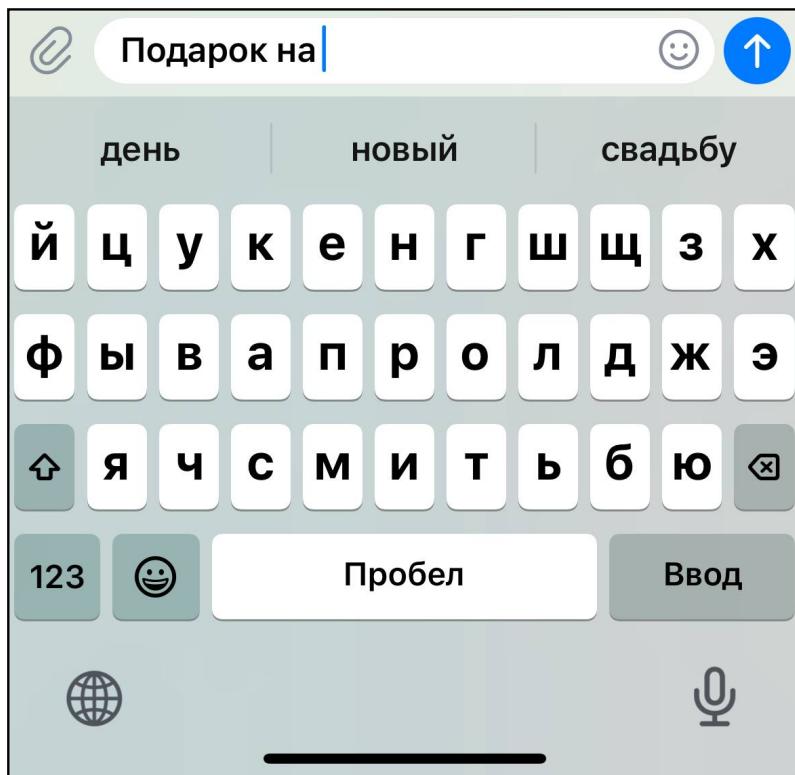
К X 🔍

- 🔍 кроссовки женские
- 🔍 кроссовки мужские
- 🔍 кроссовки мужские летние
- 🔍 кроссовки для мальчика
- 🔍 кроссовки
- 🔍 кроссовки для девочки
- 🔍 кеды женские

Примеры систем поисковых подсказок



Примеры систем поисковых подсказок



План построения системы

План построения системы

- Определяем бизнес-проблему

План построения системы

- Определяем бизнес-проблему
- Определяем метрики оценки качества работы системы

План построения системы

- Определяем бизнес-проблему
- Определяем метрики оценки качества работы системы
- Кратко обрисовываем пайплайн системы

План построения системы

- Определяем бизнес-проблему
- Определяем метрики оценки качества работы системы
- Кратко обрисовываем пайплайн системы
- Определяем необходимые данные

План построения системы

- Определяем бизнес-проблему
- Определяем метрики оценки качества работы системы
- Кратко обрисовываем пайплайн системы
- Определяем необходимые данные
- Подробно прорабатываем каждую часть пайплайна

Бизнес-проблема

Бизнес-проблема

С точки зрения пользователя

Бизнес-проблема

С точки зрения пользователя

- Пользователи не всегда знают, **что они хотят найти**
- Пользователи не всегда знают, **как называется** то, что они хотят найти
- Пользователи не всегда знают, **как сформулировать** свой запрос, чтобы найти то, что им нужно
- Пользователи допускают **опечатки** при вводе поискового запроса

Бизнес-проблема

С точки зрения системы

Бизнес-проблема

С точки зрения системы

- Помочь пользователю сформулировать свои потребности в виде текстового запроса
- Привести пользователя в более удачный (полезный) запрос с точки зрения товарной выдачи
- Сократить пользовательский путь до покупки нужного товара

Метрики оценки качества

Метрики оценки качества

- Интерактив с поисковыми подсказками:
 - CTR подсказок

Метрики оценки качества

- Интерактив с поисковыми подсказками:
 - CTR подсказок
- Качество товарной выдачи:
 - Интерактив с товарами после клика на подсказку
 - Количество заказов и GMV после клика на подсказку

Метрики оценки качества

- Интерактив с поисковыми подсказками:
 - CTR подсказок
- Качество товарной выдачи:
 - Интерактив с товарами после клика на подсказку
 - Количество заказов и GMV после клика на подсказку
- Количество опечаток в пользовательских запросах

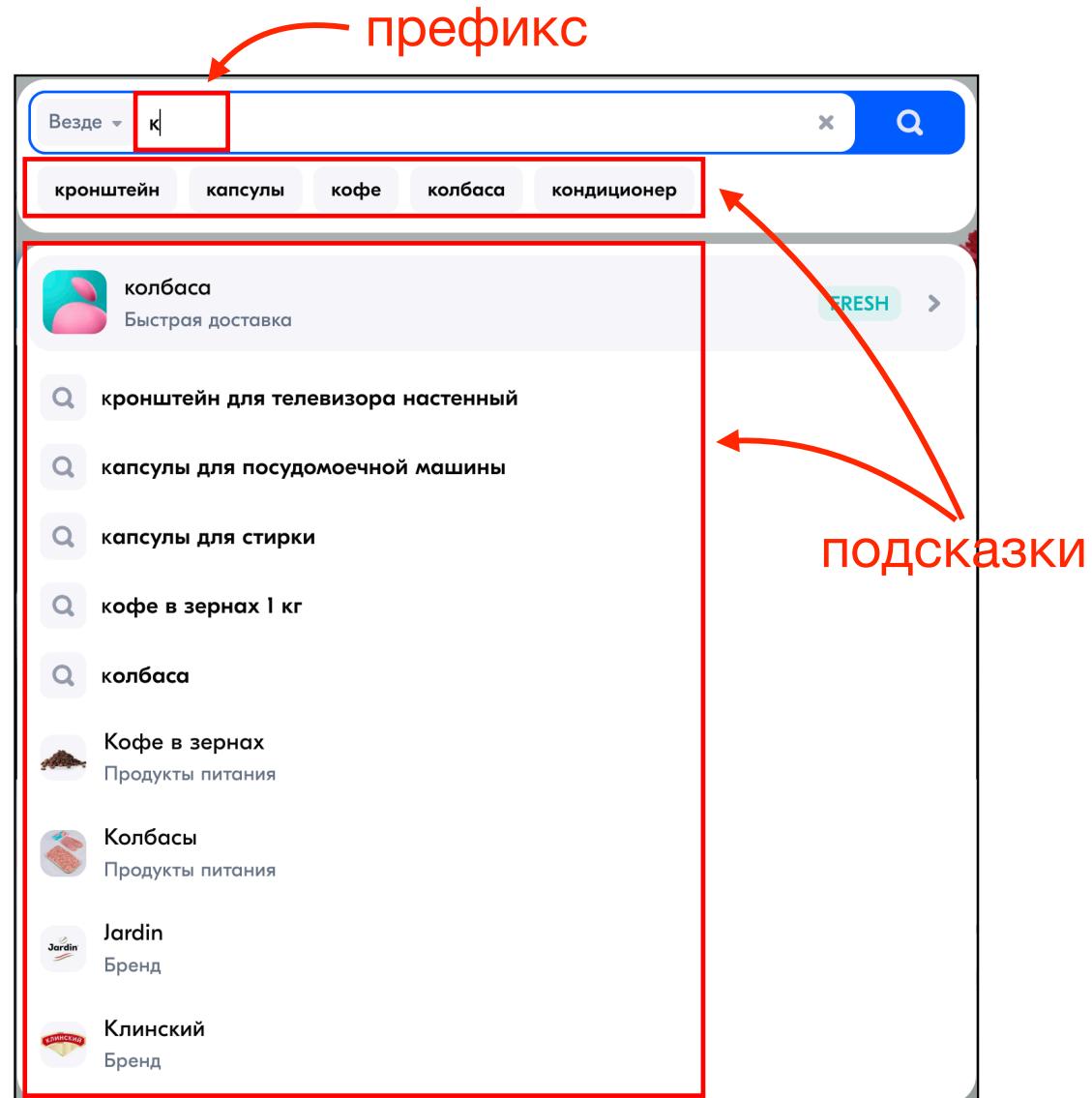
Метрики оценки качества

- Интерактив с поисковыми подсказками:
 - CTR подсказок
- Качество товарной выдачи:
 - Интерактив с товарами после клика на подсказку
 - Количество заказов и GMV после клика на подсказку
- Количество опечаток в пользовательских запросах
- Временные метрики:
 - Скорость ввода запроса
 - Время от начала ввода запроса до покупки товара

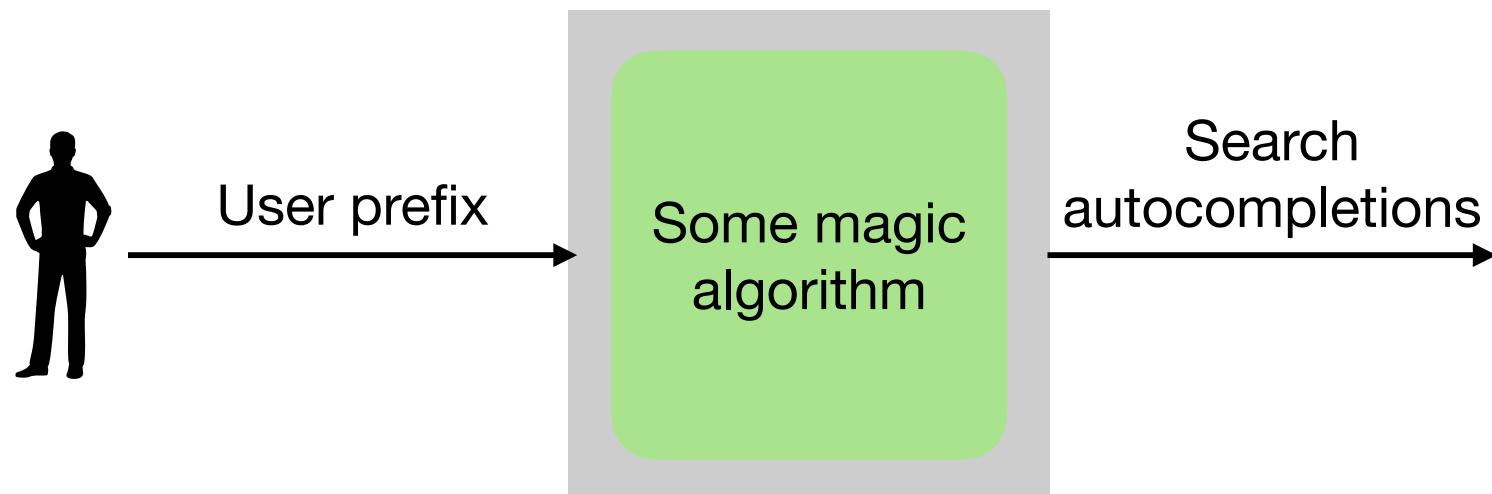
Пайпайн работы системы

Терминология

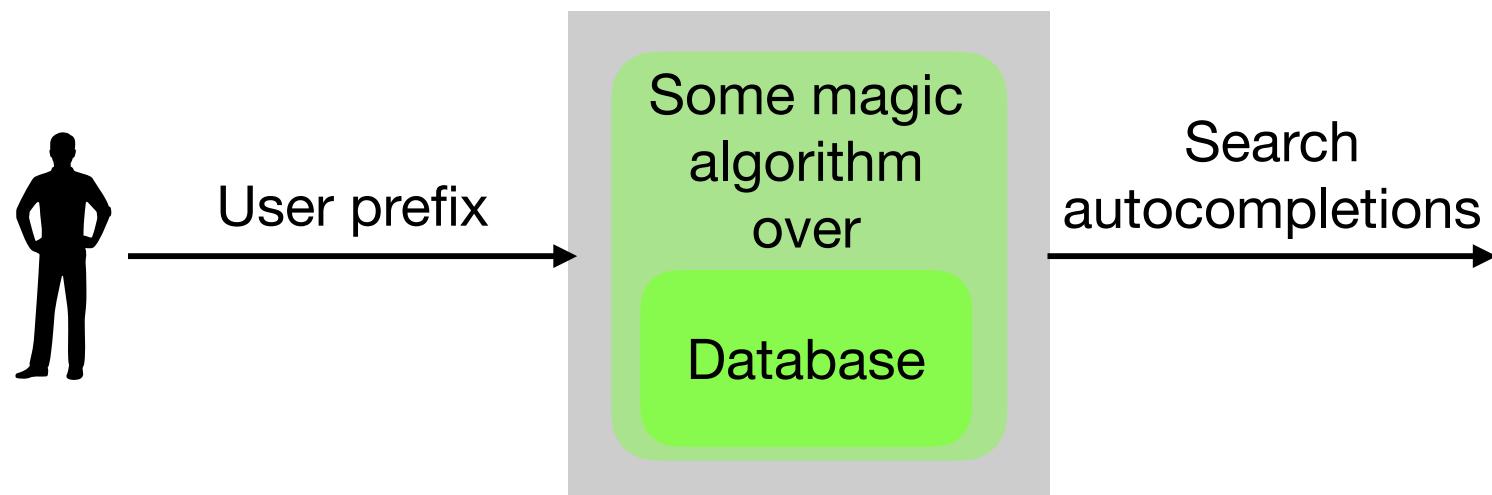
- **Префикс** — текст, который вводит пользователь в поисковую строку
- **Подсказка (сайджест, search autocomplete)** — текст, который мы предлагаем пользователю в качестве продолжения префикса



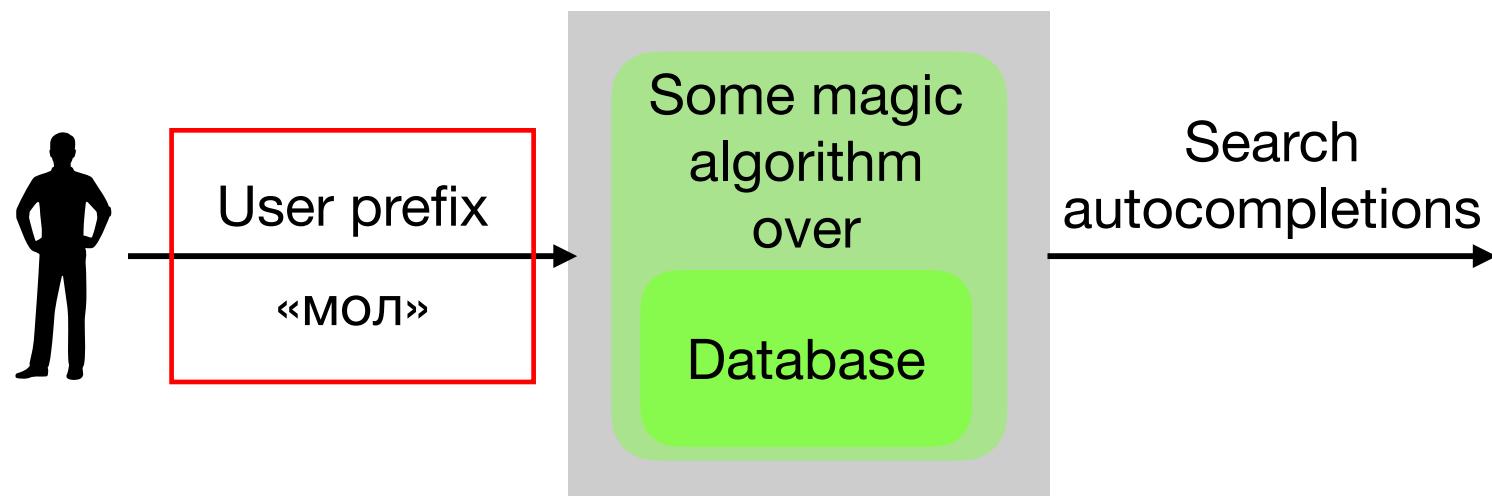
Система поисковых подсказок



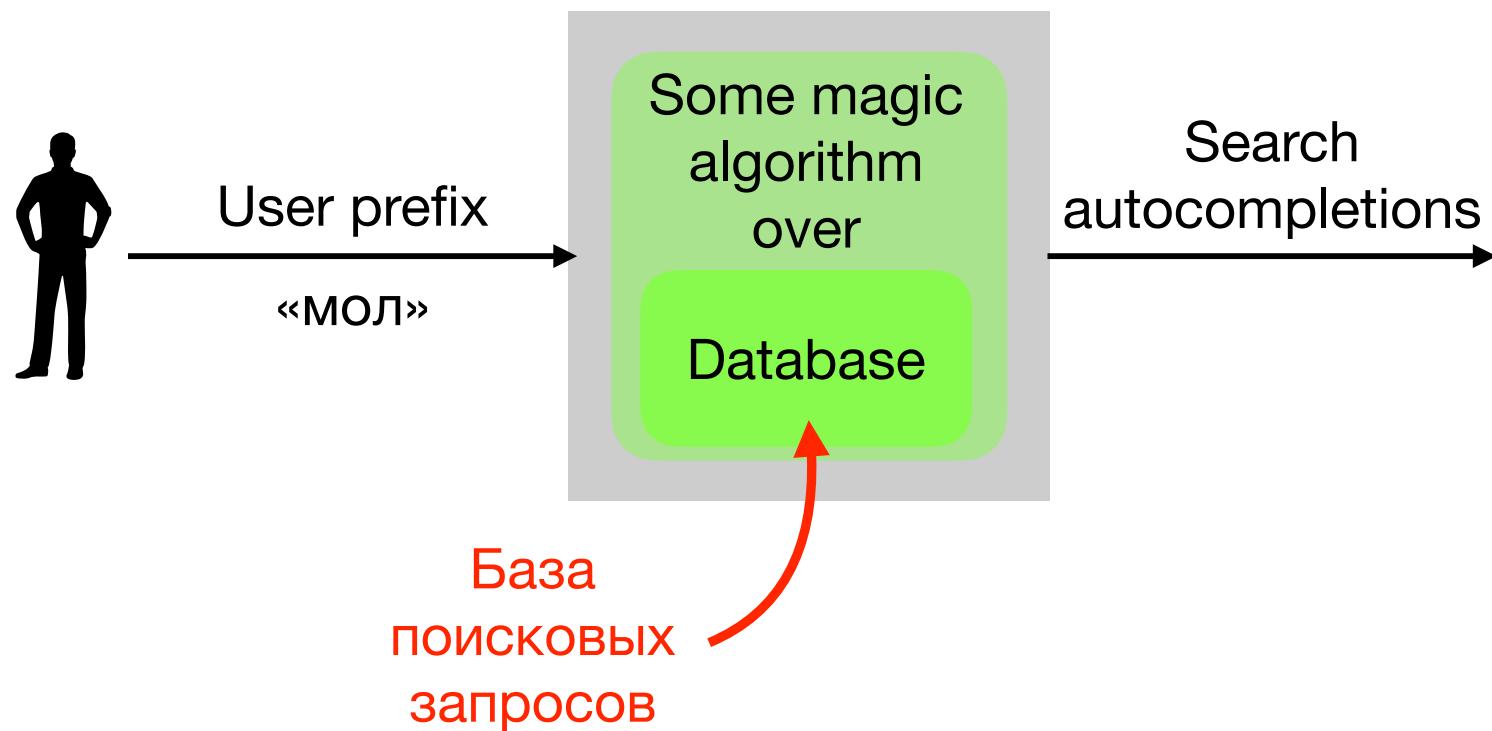
Система поисковых подсказок



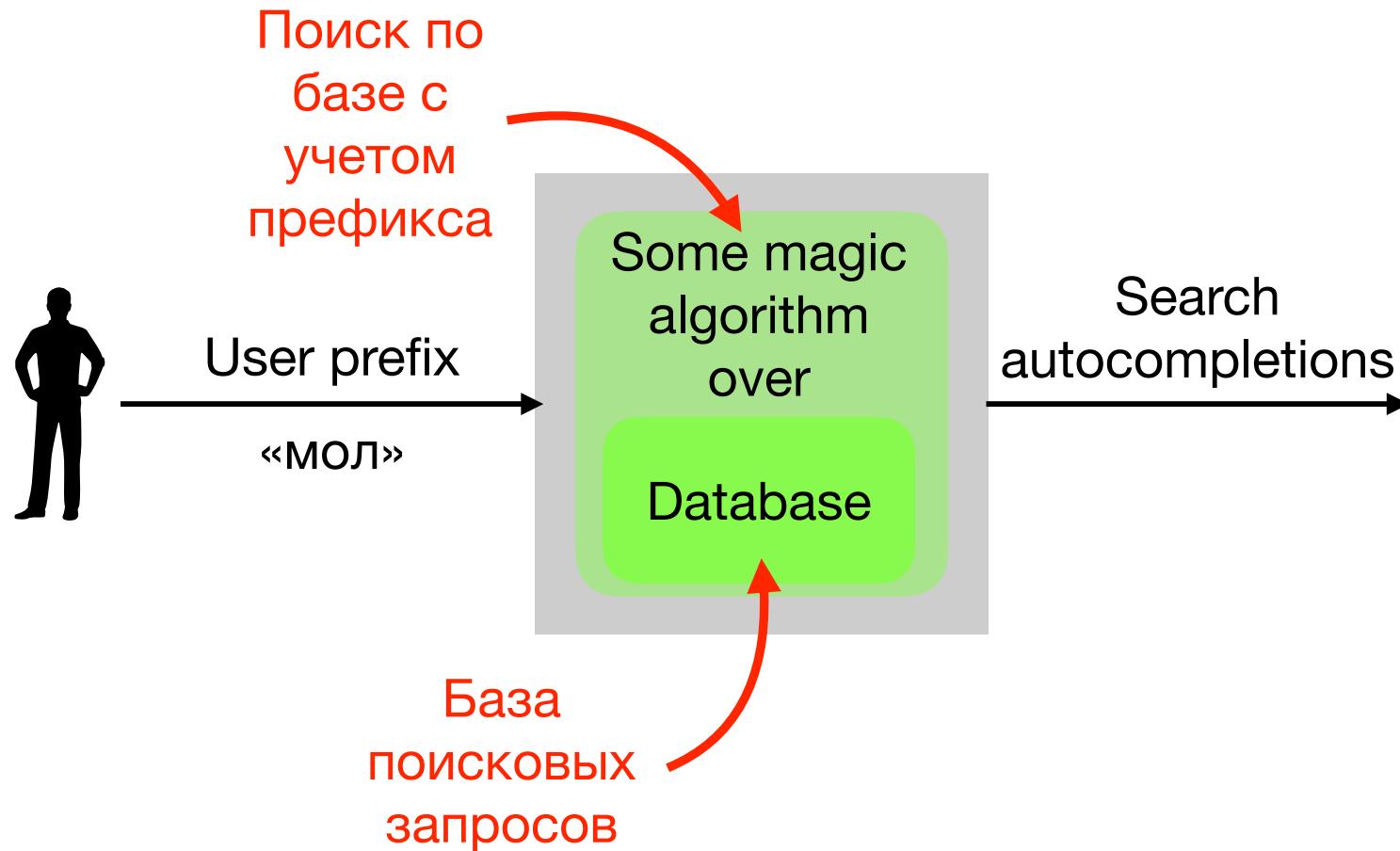
Система поисковых подсказок



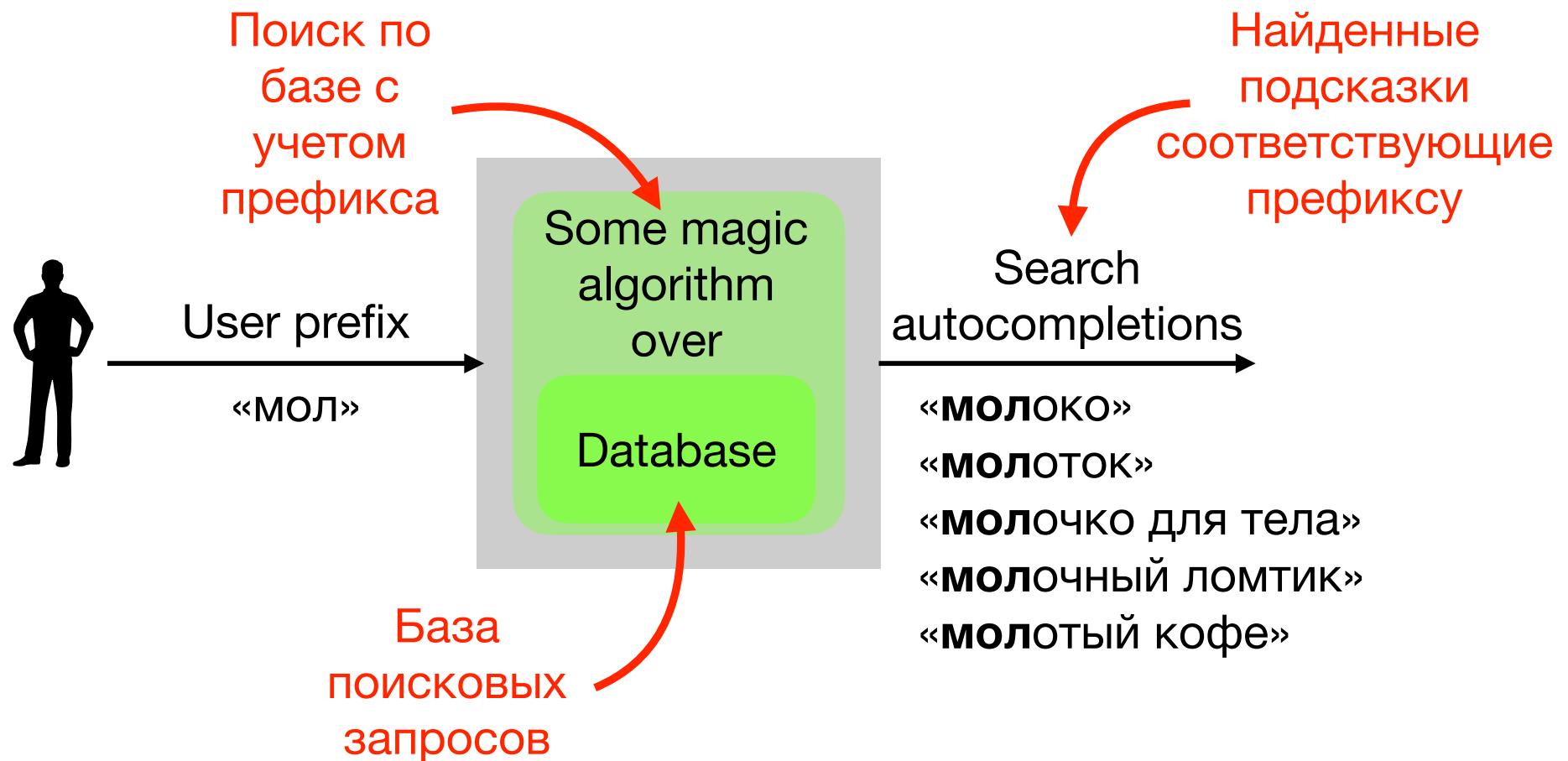
Система поисковых подсказок



Система поисковых подсказок



Система поисковых подсказок



Система поисковых подсказок

- **База** поисковых запросов
- Алгоритм **префиксного** поиска по базе запросов

Данные

Данные

- У нас есть информация:

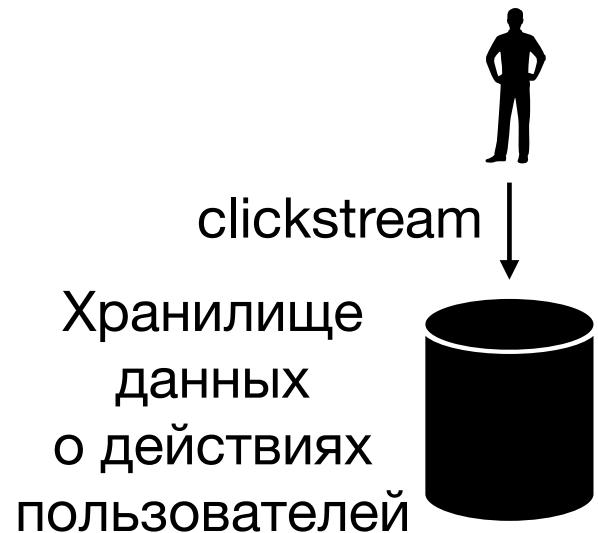
Данные

- У нас есть информация:
 - Какие запросы вводили пользователи
 - Какие запросы были полезны
 - Какие запросы привели пользователей к покупкам

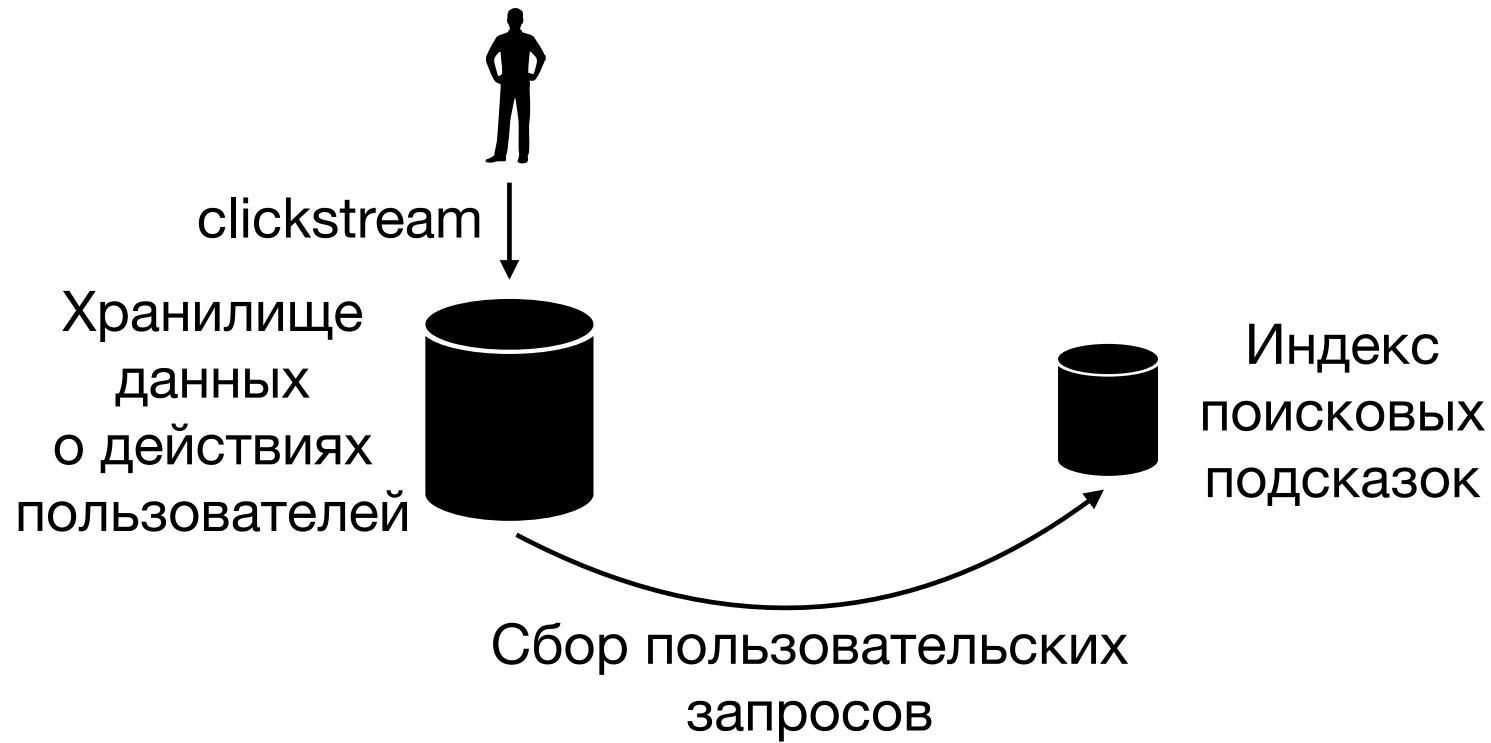
Данные

- У нас есть информация:
 - Какие запросы вводили пользователи
 - Какие запросы были полезны
 - Какие запросы привели пользователей к покупкам
- Давайте будем собирать агрегированную информацию о прошлых запросах пользователей и показывать подсказки из этого набора!

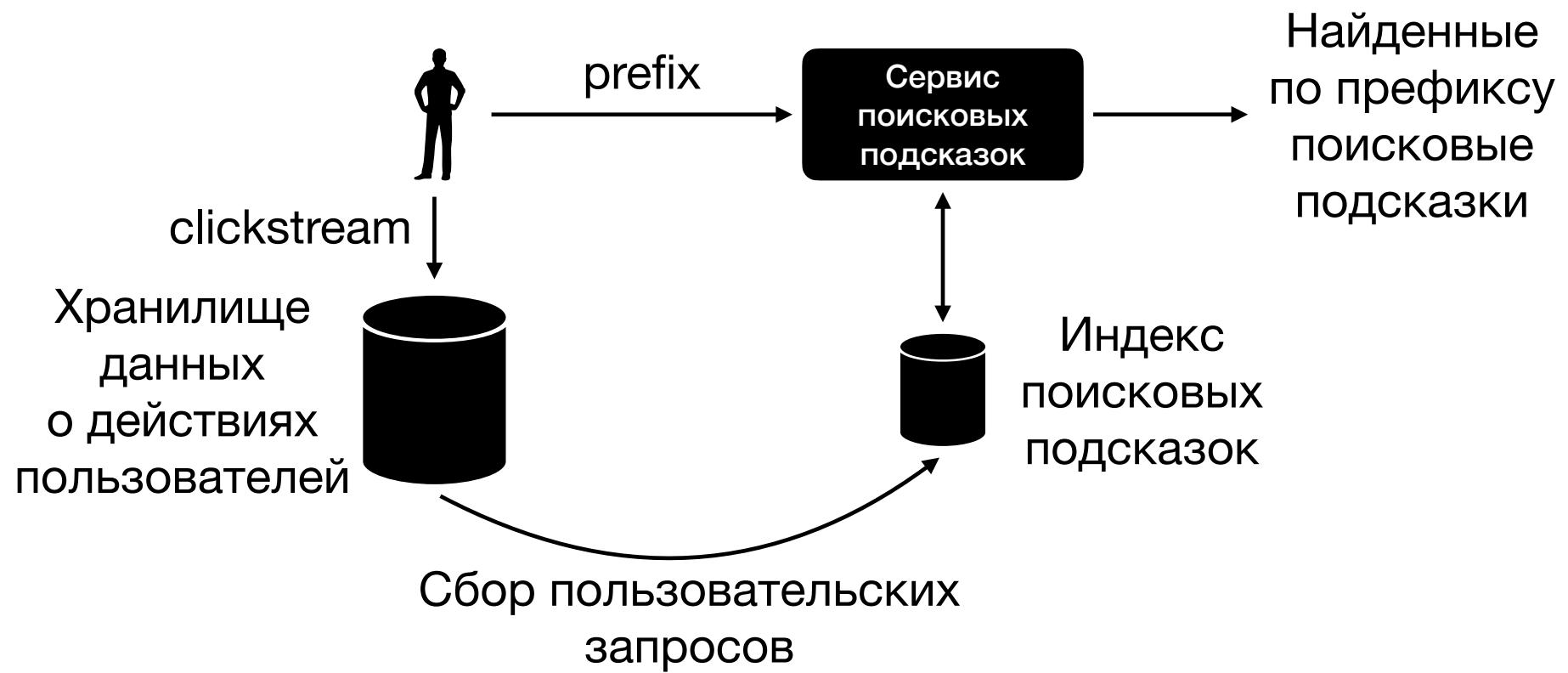
Система поисковых подсказок



Система поисковых подсказок

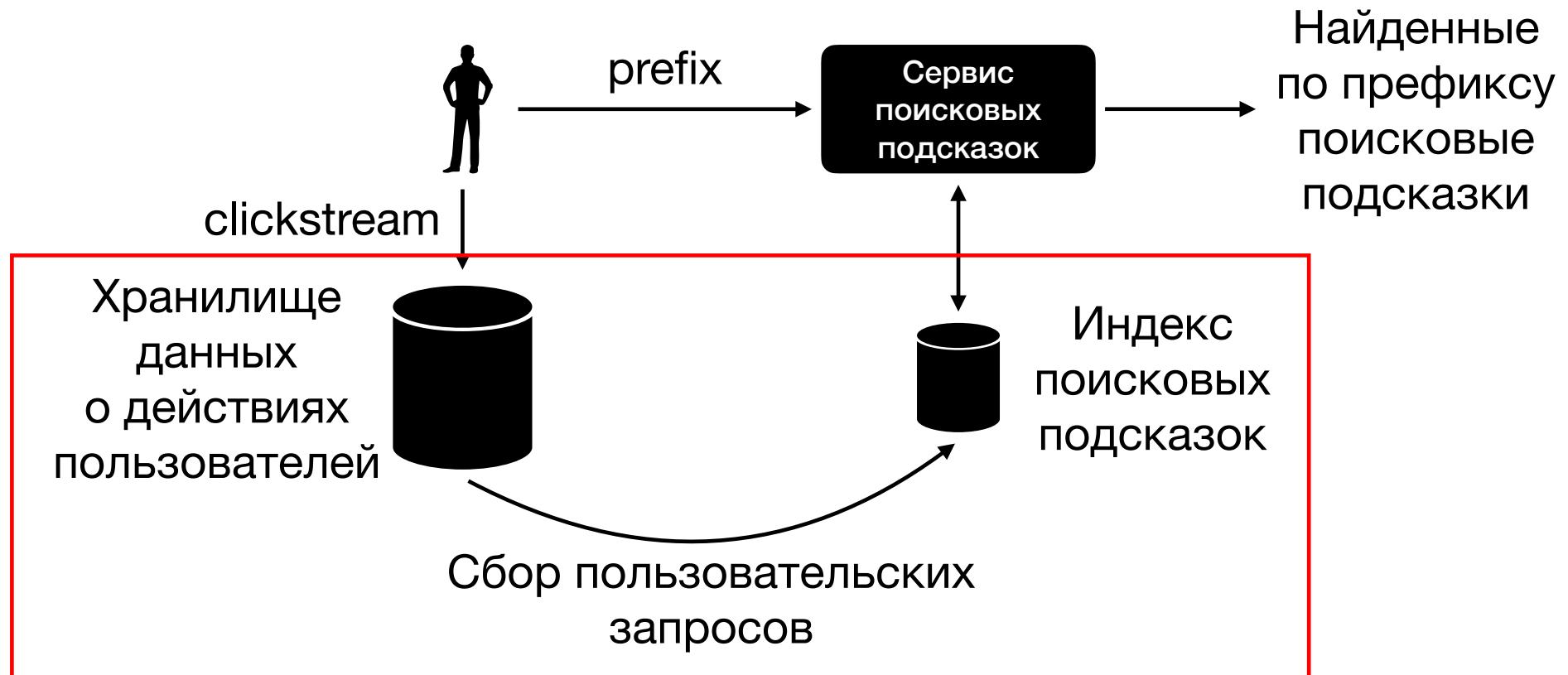


Система поисковых подсказок



Пайпайн – подробно

Сбор пользовательских запросов



Сбор пользовательских запросов

- Это может быть джоба, запускающаяся по некоторому расписанию, чтобы данные о запросах были актуальными
- Может выполняться различными *big-data* инструментами:
 - SQL / Spark / ...

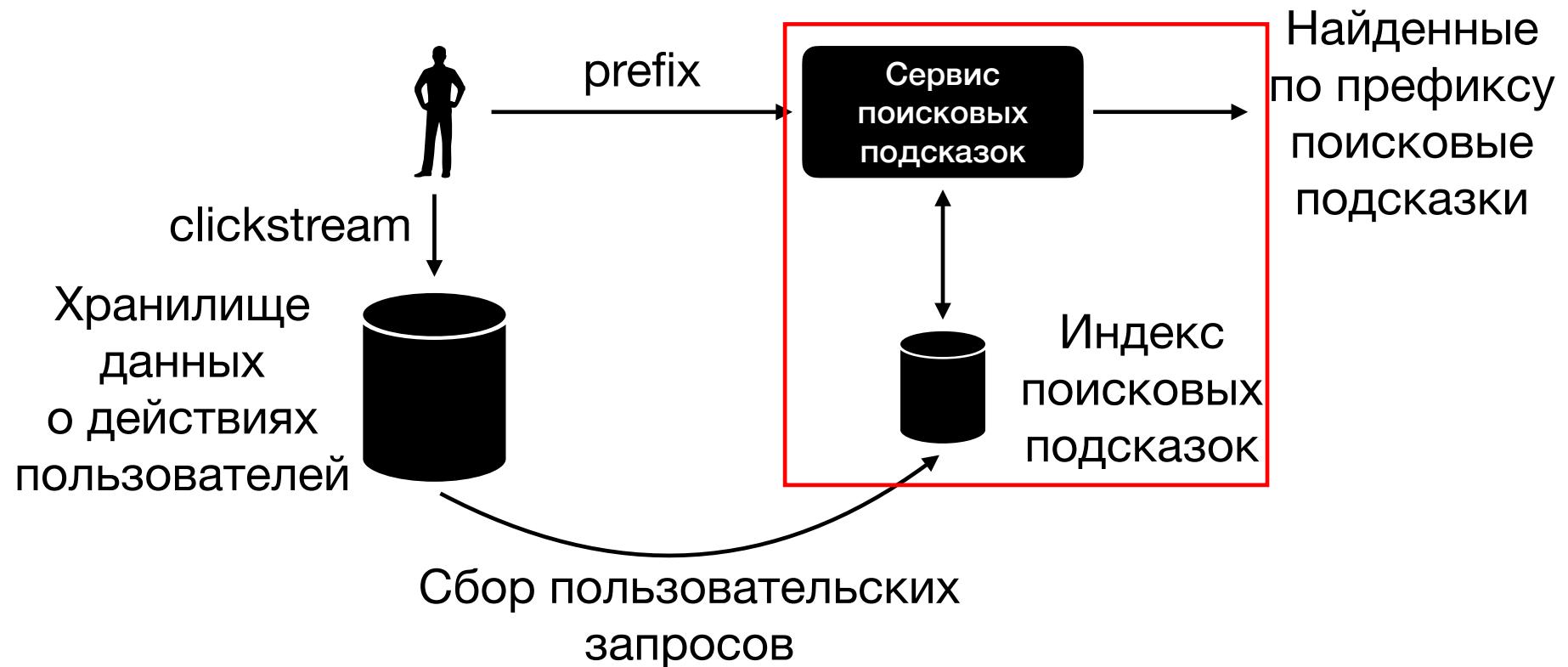
Сбор пользовательских запросов

- Это может быть джоба, запускающаяся по некоторому расписанию, чтобы данные о запросах были актуальными
- Может выполняться различными big-data инструментами:
 - SQL / Spark / ...
- Но как эффективно искать среди запросов те, которые соответствуют префиксу, введенному пользователем?

Сбор пользовательских запросов

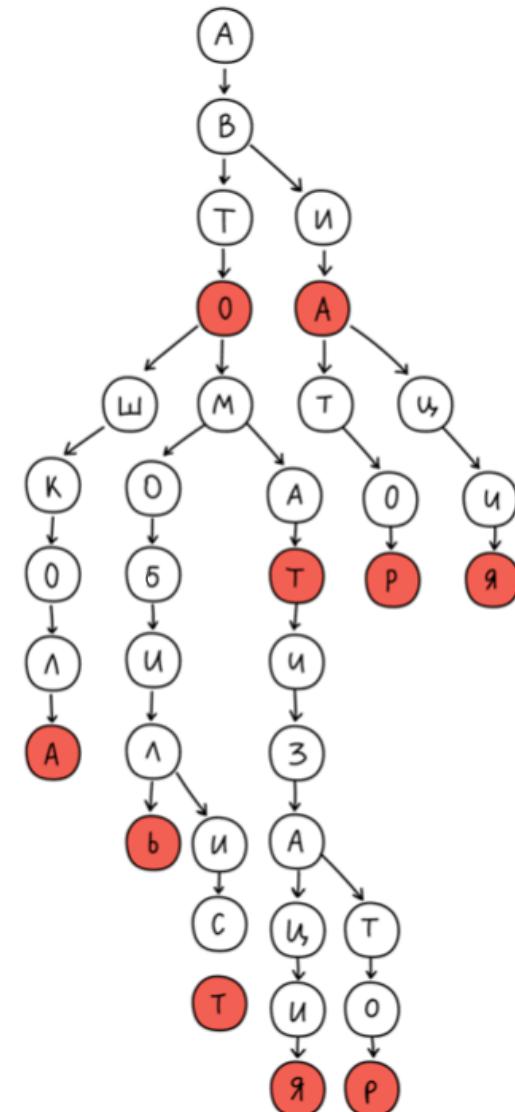
- Это может быть джоба, запускающаяся по некоторому расписанию, чтобы данные о запросах были актуальными
- Может выполняться различными big-data инструментами:
 - SQL / Spark / ...
- Но как эффективно искать среди запросов те, которые соответствуют префиксу, введенному пользователем?
 - Нужно представить данные таким образом, чтобы в онлайне осуществлять эту операцию быстро

Префиксный поиск



Префиксный поиск

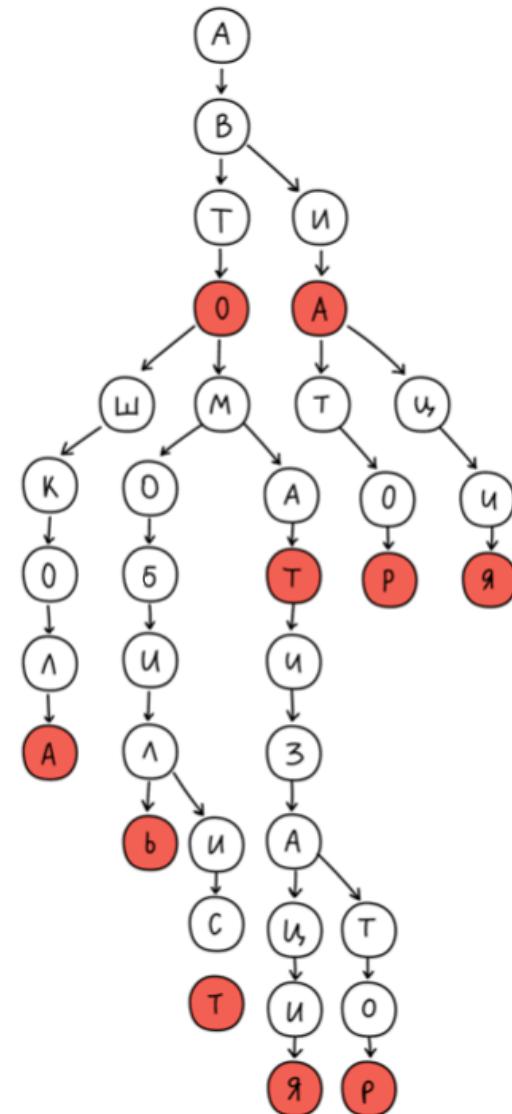
- Ключевая структура данных – префиксное дерево поиска (Trie)
- Узлами дерева являются символы
- Путь от корня к любому узлу дерева – префикс
- Красные узлы означают, что это полноценный поисковый запрос



Предфиксный поиск

Пример

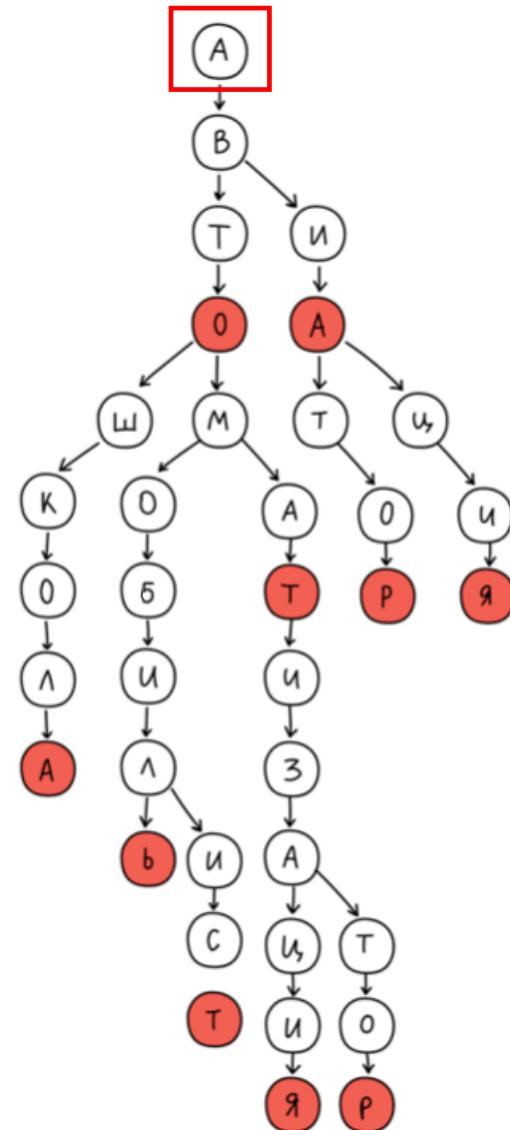
- Пользователь вводит «автом»



Предфиксный поиск

Пример

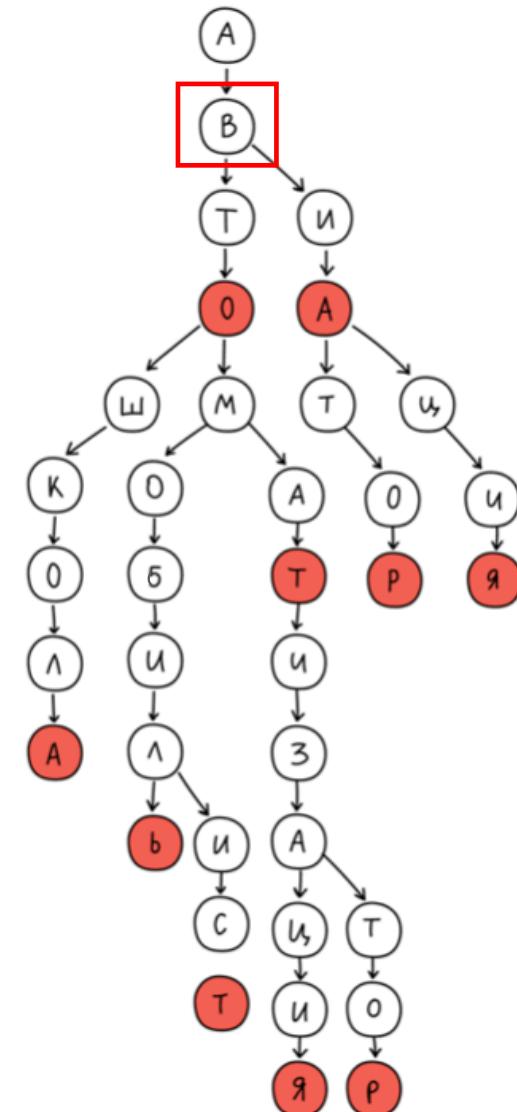
- Пользователь вводит «автом»
- Идем по префиксному дереву поиска



Префиксный поиск

Пример

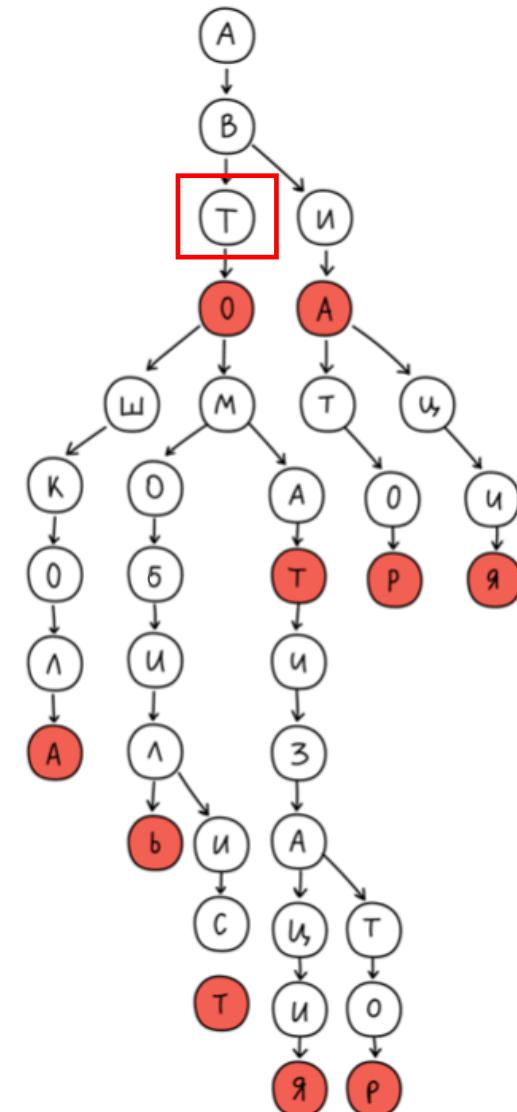
- Пользователь вводит «автом»
- Идем по префиксному дереву поиска



Предфиксный поиск

Пример

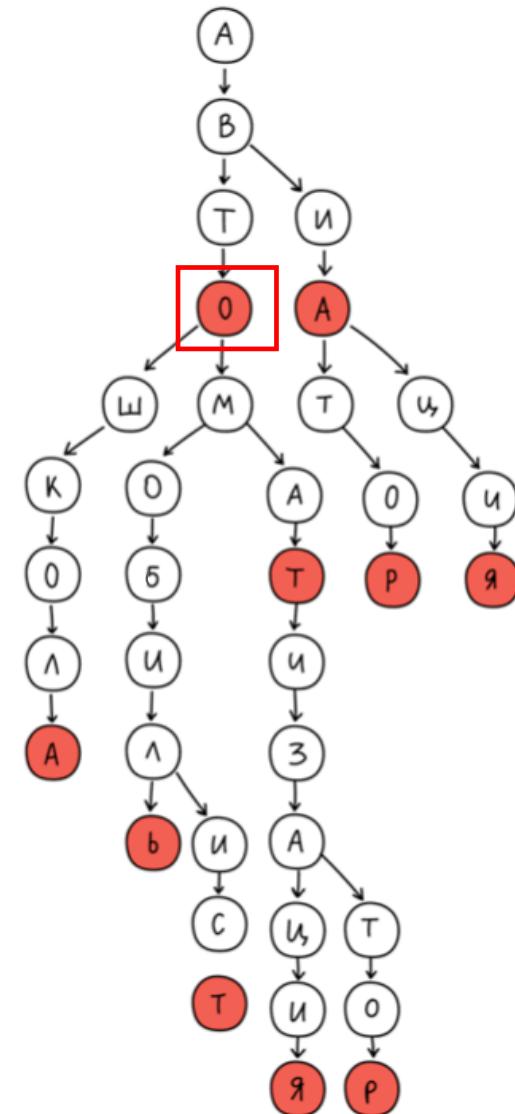
- Пользователь вводит «автом»
- Идем по префиксному дереву поиска



Предфиксный поиск

Пример

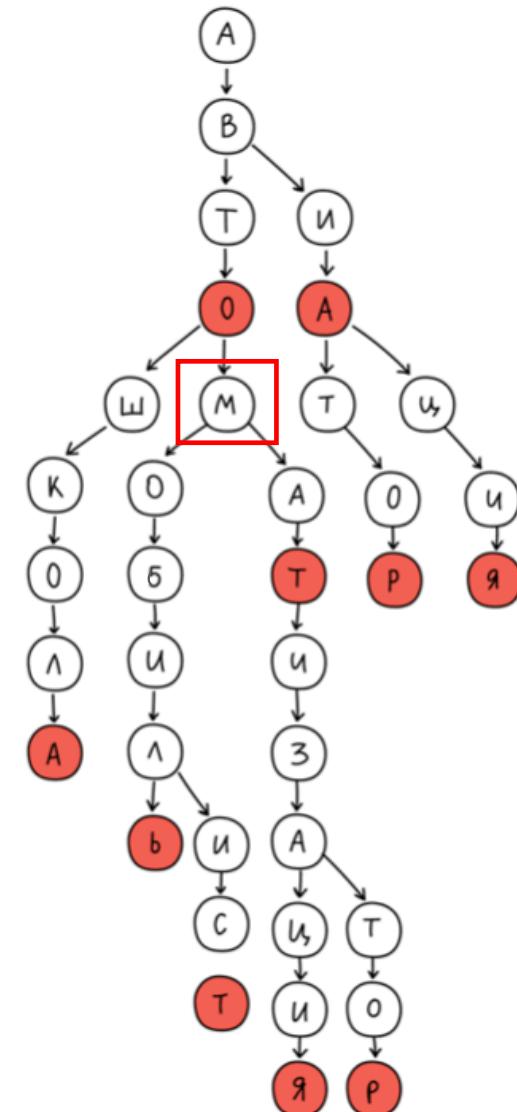
- Пользователь вводит «автом»
- Идем по префиксному дереву поиска



Предфиксный поиск

Пример

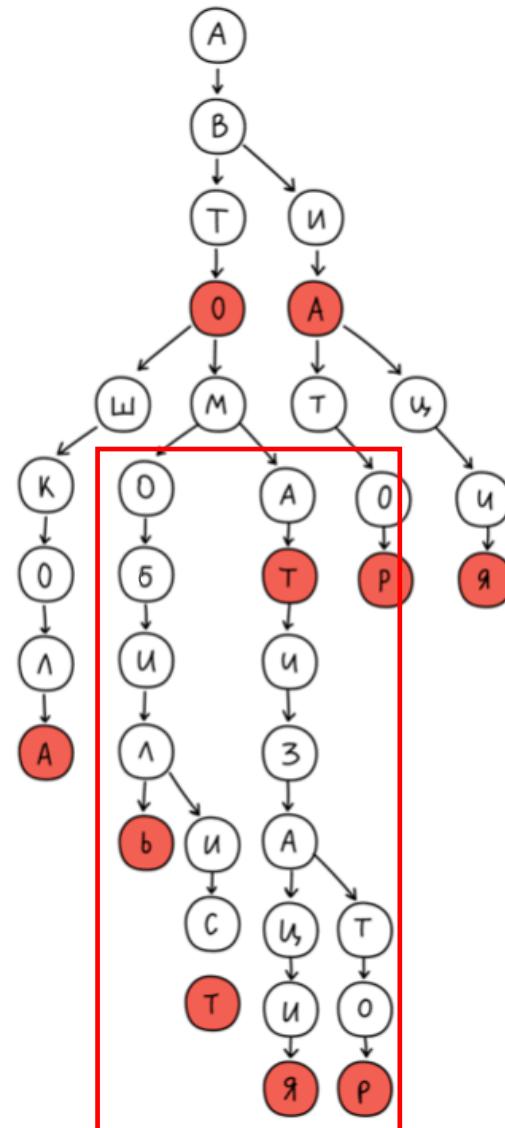
- Пользователь вводит «автом»
- Идем по префиксному дереву поиска



Префиксный поиск

Пример

- Пользователь вводит «автом»
- Идем по префиксному дереву поиска
- Получаем все запросы, соответствующие префиксу (оставшееся поддерево):
[«автомобиль», «автомобилист», «автоматизация», «автоматизатор»]



Префиксный поиск

Недостатки текущего решения

Префиксный поиск

Недостатки текущего решения

- По коротким префиксам мы будем получать тысячи подсказок
- Многие из этих подсказок очень редко используемые и нерелевантные

Префиксный поиск

Недостатки текущего решения

- По коротким префиксам мы будем получать тысячи подсказок
- Многие из этих подсказок очень редко используемые и нерелевантные
- Решение:
 - Давайте для каждого запроса хранить не только текст, но ещё и популярность / полезность запроса

Префиксный поиск

Популярность запроса

Префиксный поиск

Популярность запроса

- Количество поисков с таким запросом

Префиксный поиск

Популярность запроса

- Количество поисков с таким запросом
 - Проблема fraud-a / накрутки / ботов
 - Целью нашей системы является помочь пользователю в нахождении нужного товара

Префиксный поиск

Популярность запроса

- Количество поисков с таким запросом
 - Проблема fraud-a / накрутки / ботов
 - Целью нашей системы является помочь пользователю в нахождении нужного товара
- Количество целевых действий с товаром по такому запросу
 - Будем считать количество купленных товаров по запросам за предыдущие N дней — довольно неплохая характеристика популярности + полезности запроса

Префиксный поиск

На практике

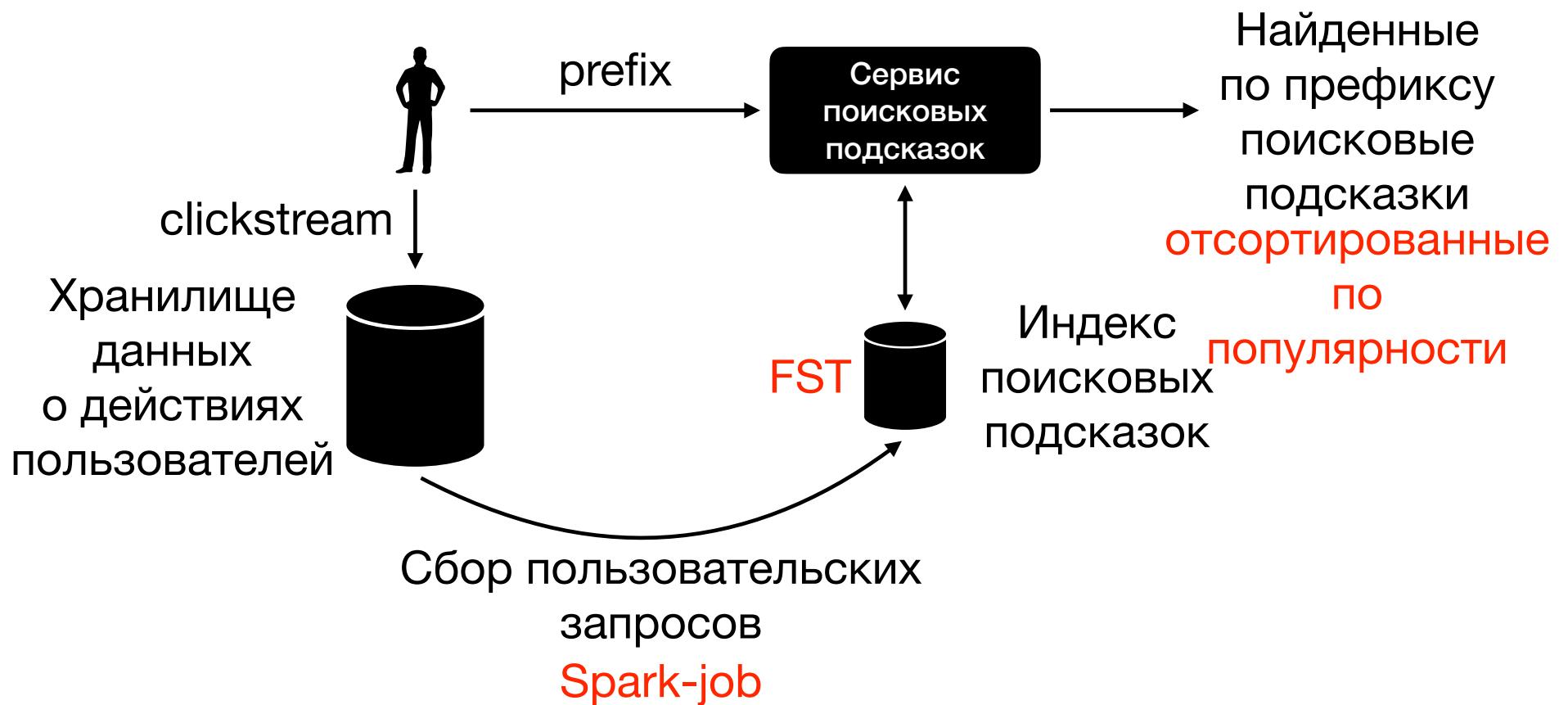
- Обычно вместо Trie используется FST — finite-state transducer
- FST представляет собой граф:
 - узлы (состояния) — символы
 - в отличие от дерева одинаковые суффиксы запросов тоже объединены в единый путь в графе
- FST позволяет:
 - хранить индекс в сжатом виде (слияние не только префиксов, но и суффиксов)
 - осуществлять поиск ещё быстрее
 - учитывать веса запросов при поиске
 - хранить метаданные запросов

Префиксный поиск

Trie vs. FST

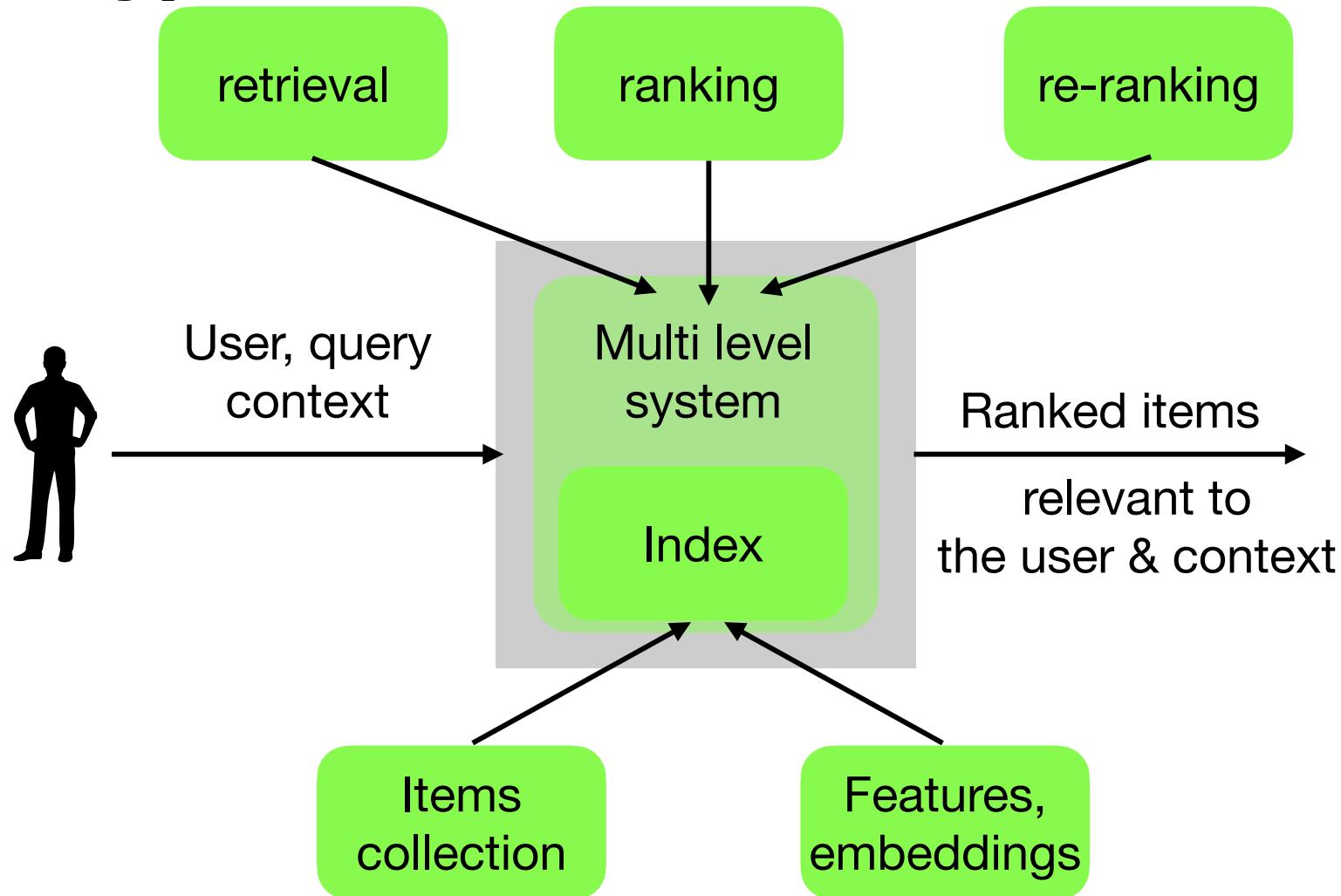
	Trie	FST
Память	Высокий расход (каждый символ — отдельный узел)	Сжатое хранение (общие префиксы/суффиксы объединены)
Скорость поиска	Быстрый ($O(L)$, где L — длина ключа)	Сопоставима с Trie, но может быть быстрее благодаря оптимизациям
Гибкость	Только ключи (без значений)	Ключи + ассоциированные значения (например, веса)
Реализация	Простая	Сложная
Поддержка операций	Вставка/удаление динамические	Часто неизменяем (оптимизирован для статических данных)

Префиксный поиск

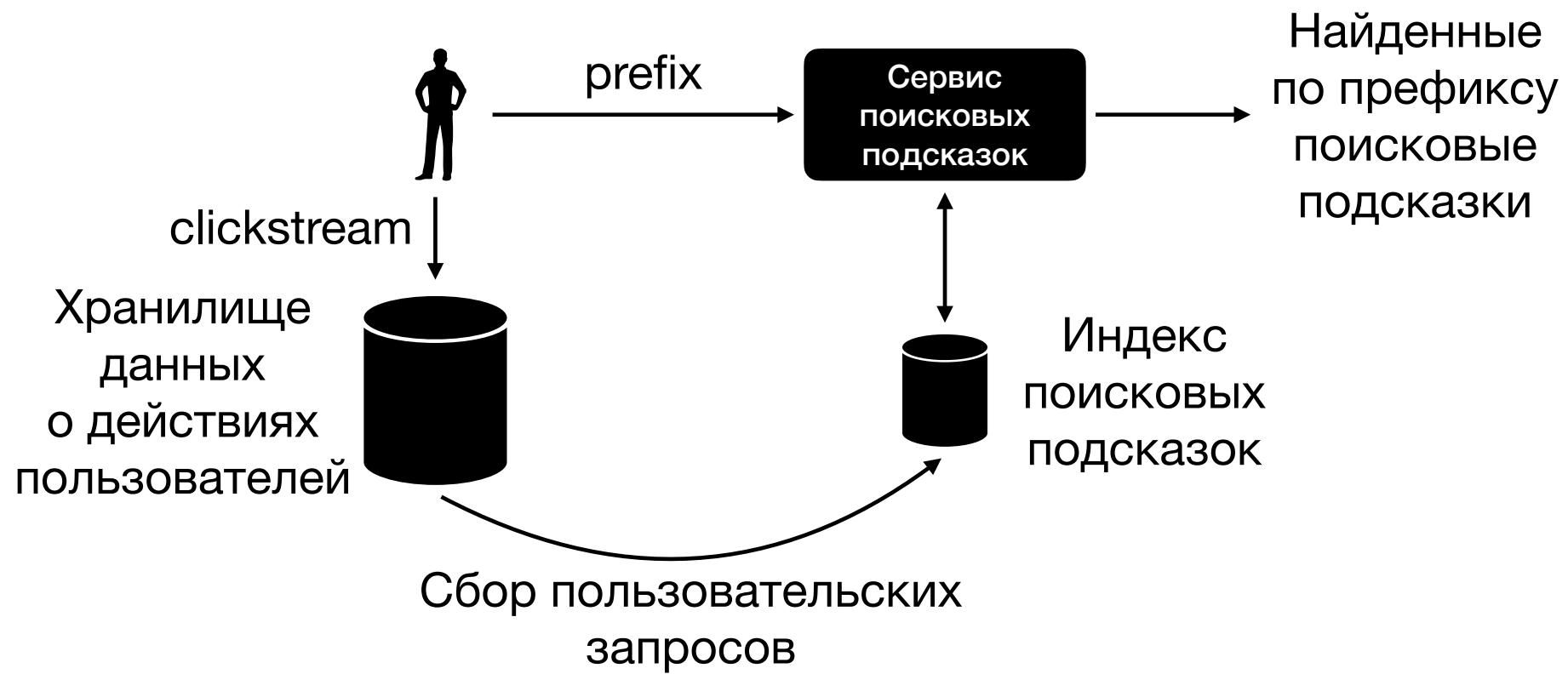


**Усложняем и улучшаем нашу
систему**

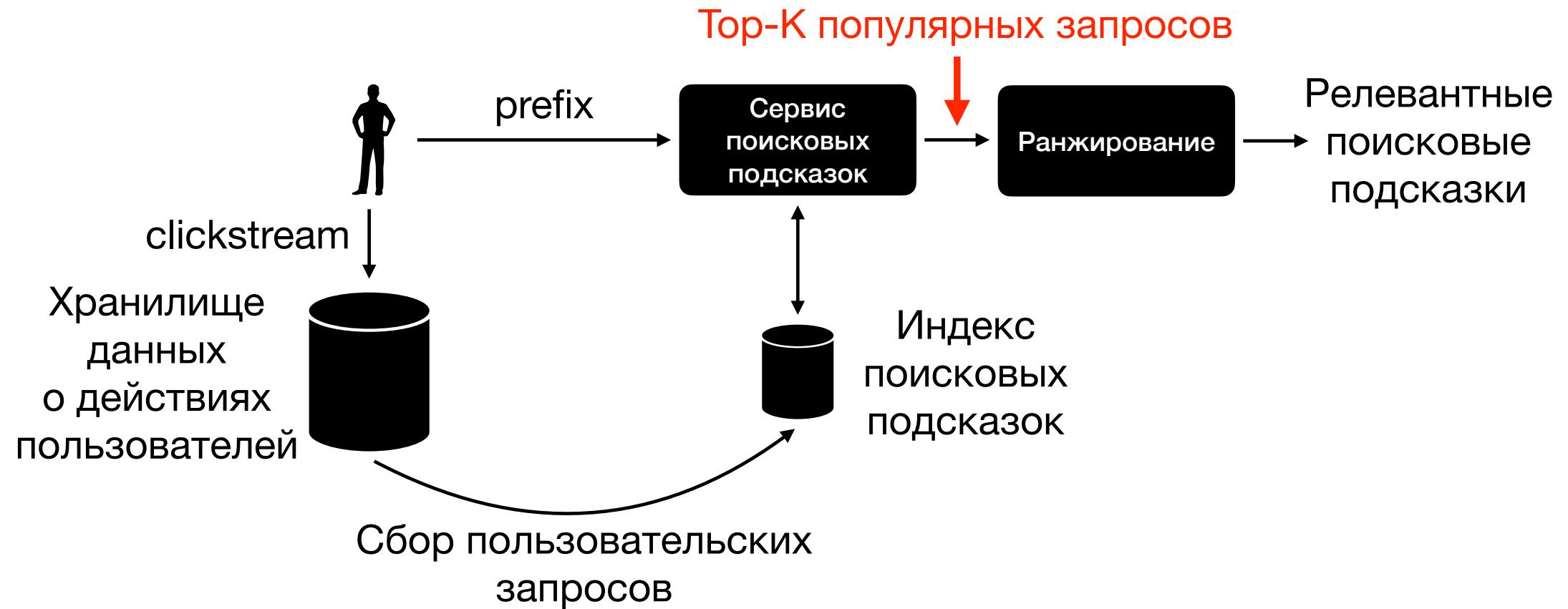
Многоуровневая система



Система подсказок



Многоуровневая система подсказок



Система поисковых подсказок

- **База** поисковых запросов
- Алгоритм **префиксного** поиска по базе запросов
- **Ранжирование** найденных префиксным поиском подсказок

ML-постановка задачи

- X – множество запросов (подсказок)
- $f_k(x) = f_k(x \mid u = u_c)$ – признак запроса x для пользователя u_c
- y – релевантность подсказки, $y \in R$
- Пусть $i < j$ – правильный порядок на парах (i, j) , т.е. $y_i < y_j$
- **Задача** – найти $a(x)$: если $i < j$, то $a(x_i) < a(x_j)$

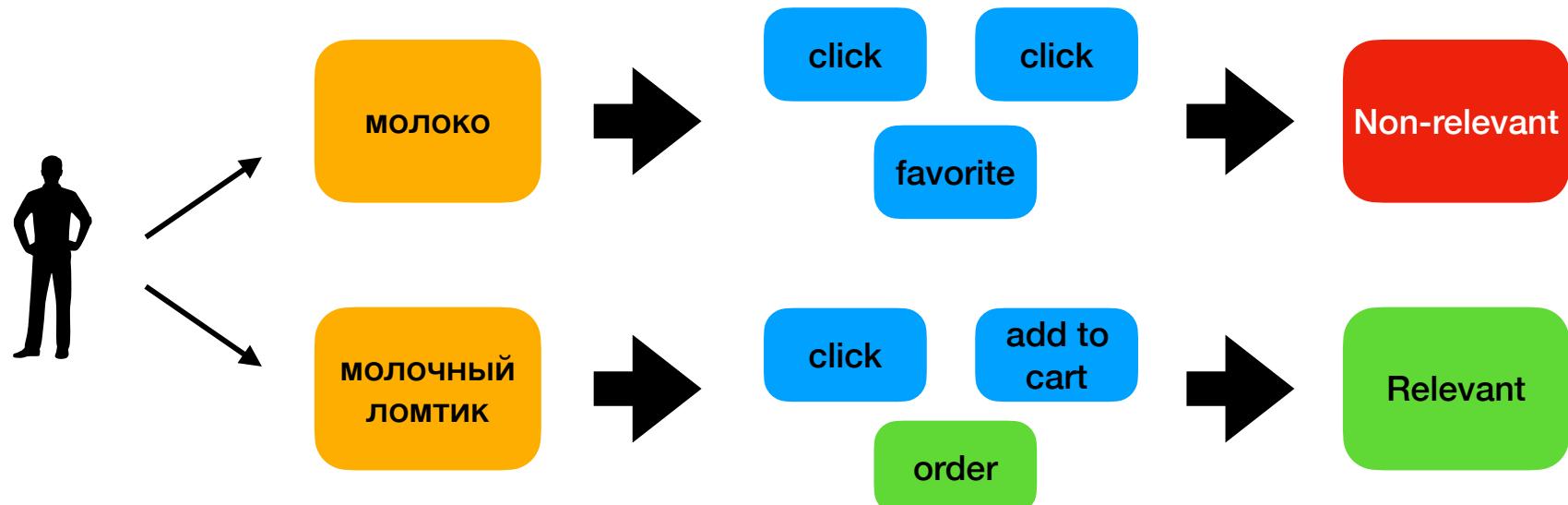
ML-постановка задачи

Группа для ранжирования	Саджест	f_1	f_2	...	f_m	Релевантность
user_id = 123 prefix = "M"	«МОЛОКО»	342	0.05	...	0.78	1
	«макароны»	435	0.91	...	0.09	3
	«мыло для рук»	121	1.02	...	0.61	2
	«МОЛОТОК»	97	2.5	...	0.34	0

Релевантность подсказки пользователю

Релевантность подсказки пользователю

- Основная цель — помочь пользователю найти нужный товар
- Будем считать запрос релевантным, если пользователь после ввода данного запроса заказал товар



Признаковое описание подсказок

Признаковое описание подсказок

- Статистические признаки запроса
 - описывают популярность запроса

Признаковое описание подсказок

- Статистические признаки запроса
 - описывают популярность запроса
- Персональные признаки пользователя
 - говорят о том, что перед нами за человек — описывают long-term предпочтения пользователя

Признаковое описание подсказок

- Статистические признаки запроса
 - описывают популярность запроса
- Персональные признаки пользователя
 - говорят о том, что перед нами за человек — описывают long-term предпочтения пользователя
- Контекстные признаки
 - опираются на предыдущую активность пользователя в сессии — описывают short-term предпочтения пользователя

Признаковое описание подсказок

Статистические признаки запроса

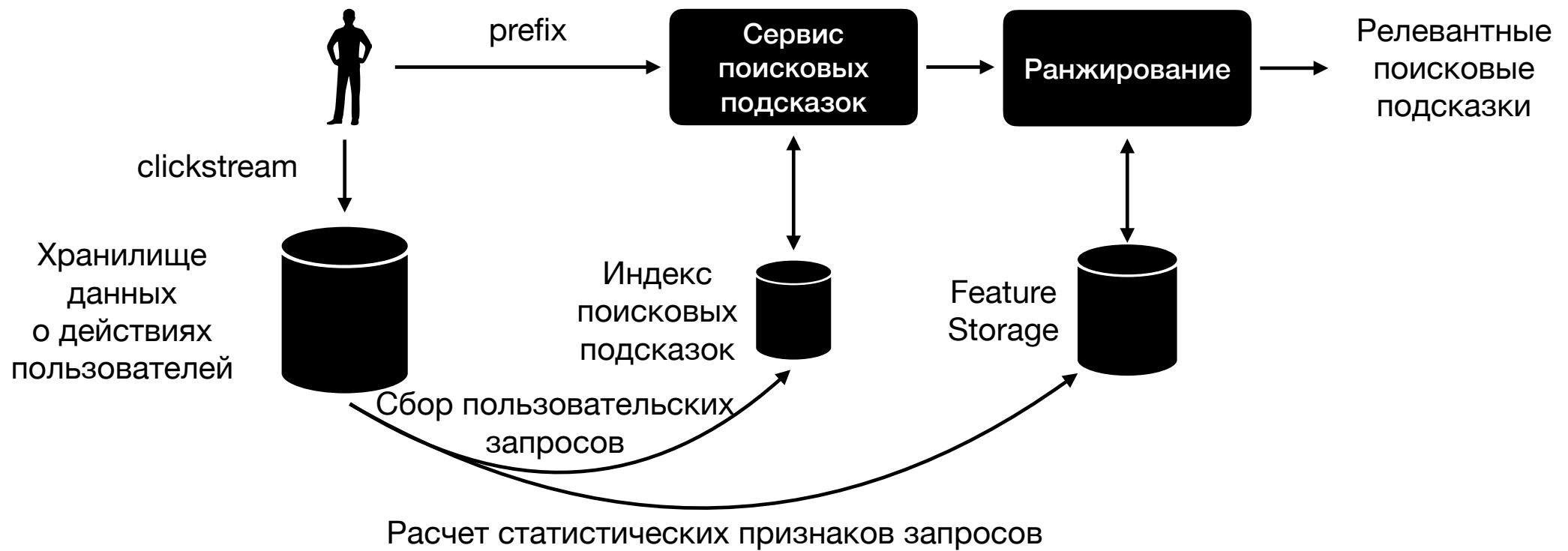
- Счетчики количества полезных событий по запросу
 - кол-во добавлений товаров в корзину / в избранное по запросу
 - кол-во заказов товаров по запросу
 - ...
- Запросные конверсии
 - конверсия из просмотра товара в добавление к корзину
 - ...
- Фичи сезонности запросов
 - популярность запроса в мае

Признаковое описание подсказок

Статистические признаки запроса

- Преимущества:
 - Дают много информации о популярности запросов
 - Легко считаются любым ETL-процессом (Spark, ClickHouse, etc.)
- Недостатки:
 - Ненормированность \implies со временем распределение признаков сдвигается
 - fraud / накрутка / боты
 - Никак не учитывается персонализация

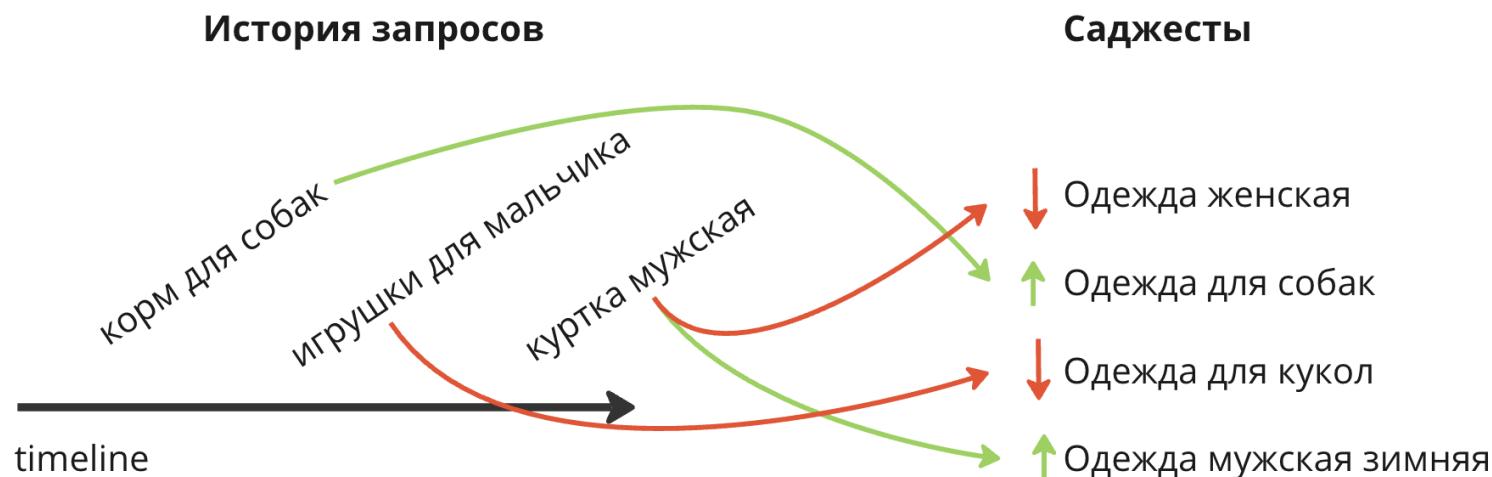
Многоуровневая система подсказок



Признаковое описание подсказок

Контекстные признаки

- **Идея:** новый поисковый запрос зачастую близок по смыслу к прошлым запросам пользователя



Признаковое описание подсказок

Контекстные признаки

- **Идея:** новый поисковый запрос зачастую близок по смыслу к прошлым запросам пользователя
- Как учитывать близость между запросом в истории и подсказкой?

Признаковое описание подсказок

Контекстные признаки

- **Идея:** новый поисковый запрос зачастую близок по смыслу к прошлым запросам пользователя
- Как учитывать близость между запросом в истории и подсказкой?

Мера Жаккара

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

Признаковое описание подсказок

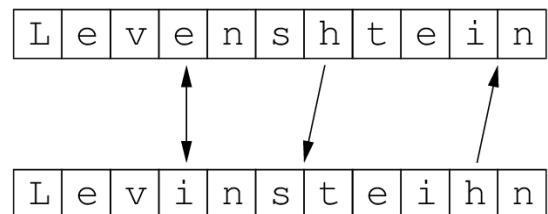
Контекстные признаки

- **Идея:** новый поисковый запрос зачастую близок по смыслу к прошлым запросам пользователя
- Как учитывать близость между запросом в истории и подсказкой?

Мера Жаккара

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

Расстояние Левенштейна



Признаковое описание подсказок

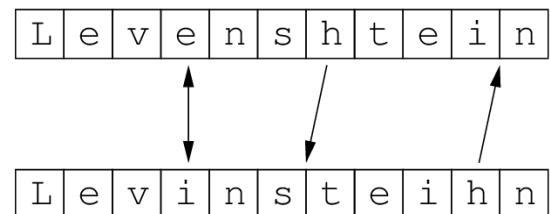
Контекстные признаки

- **Идея:** новый поисковый запрос зачастую близок по смыслу к прошлым запросам пользователя
- Как учитывать близость между запросом в истории и подсказкой?

Мера Жаккара

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

Расстояние Левенштейна



Косинусная близость

$$\cos_sim(\vec{A}, \vec{B}) = \frac{\vec{A} \cdot \vec{B}}{||\vec{A}|| \cdot ||\vec{B}||}$$

Признаковое описание подсказок

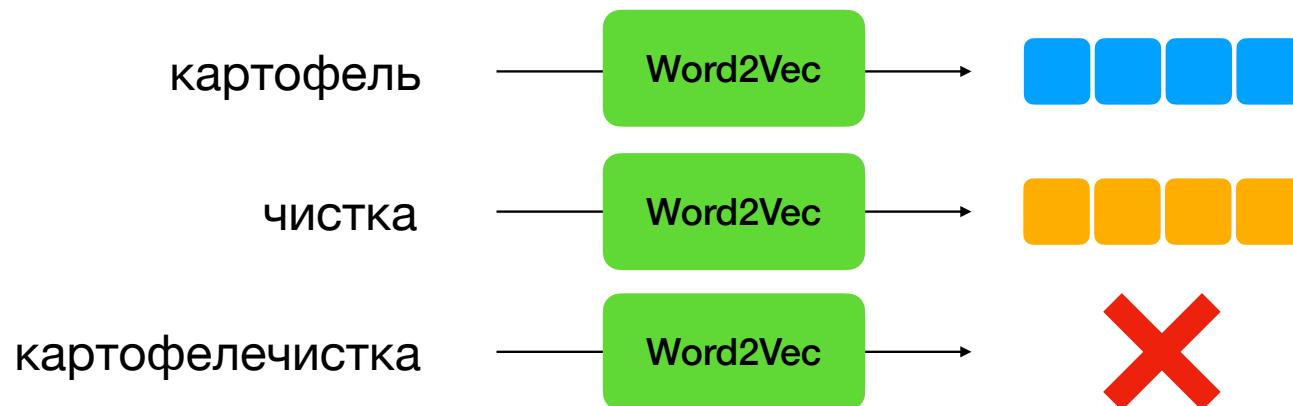
Семантическая близость запросов

- Word2Vec

Признаковое описание подсказок

Семантическая близость запросов

- Word2Vec – проблема OOV токенов



Признаковое описание подсказок

Семантическая близость запросов

- **Word2Vec** – проблема OOV токенов
- **FastText** – избегаем OOV и проблему морфологии

картофелечистка



[«<картофе», «картофел», «артофеле»,
«ртофелеч», «тофелечи», «офелечис», «фелечист»,
«елечистк», «лечистка», «ечистка»»]

Признаковое описание подсказок

Семантическая близость запросов

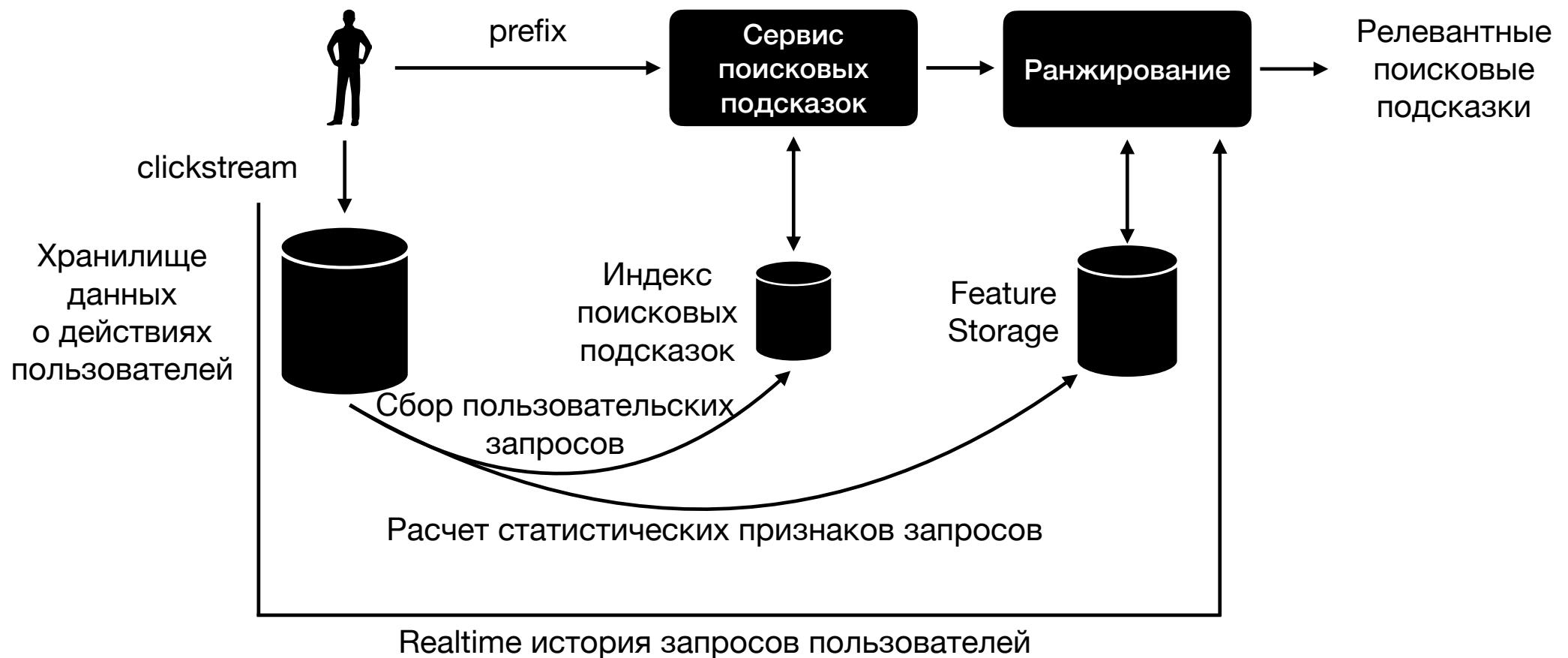
- **Word2Vec** – проблема OOV токенов
- **FastText** – избегаем OOV и проблему морфологии
- **BERT** – тяжело инференсить в онлайне (но можно делать расчет векторов асинхронно и кешировать векторы популярных запросов)

Признаковое описание подсказок

Семантическая близость запросов

- **Word2Vec** – проблема OOV токенов
- **FastText** – избегаем OOV и проблему морфологии
- **BERT** – тяжело инференсить в онлайне (но можно делать расчет векторов асинхронно и кешировать векторы популярных запросов)
- ... (любая другая языковая модель)

Многоуровневая система подсказок



Признаковое описание подсказок

Контекстные признаки

- Контекстными признаками может быть не только история поисковых запросов пользователей
- Можно также учитывать:
 - долгосрочную популярность запроса для конкретного пользователя
 - историю добавлений товаров в корзину / покупок товаров
 - историю посещений категорий в каталоге товаров
 - историю взаимодействий с товарами в рекомендациях
 - ...

Признаковое описание подсказок

Контекстные / долгосрочные признаки

- **Идея:** можно учитывать не только семантическую близость, но и близость по «интересам»

Признаковое описание подсказок

Контекстные / долгосрочные признаки

- **Идея:** можно учитывать не только семантическую близость, но и близость по «интересам»
- **ALS** на матрице взаимодействий пользователь-запрос:

R ($m \times n$)	SCOPE	FOOTBALL	CAT	MUSIC	THEATER
USER 1	+	+	-	-	-
USER 2	-	-	+	-	-
USER 3	-	+	-	+	-
USER 4	+	-	-	+	+
USER 5	-	-	-	+	+

$$P(m \times d)$$



$$Q(d \times n)$$

X



Признаковое описание подсказок

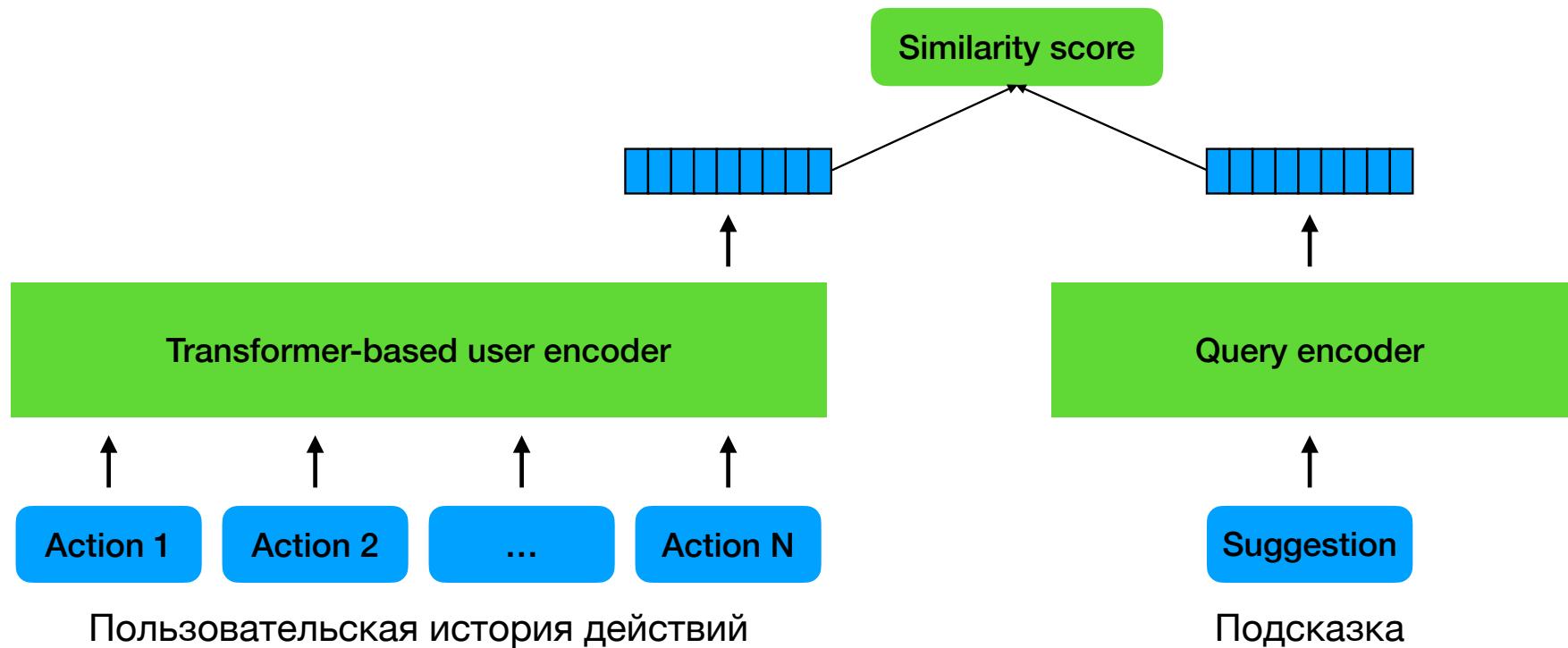
Нейросетевая персонализация

- Можно использовать те же идеи, что и в нейросетевых рекомендациях

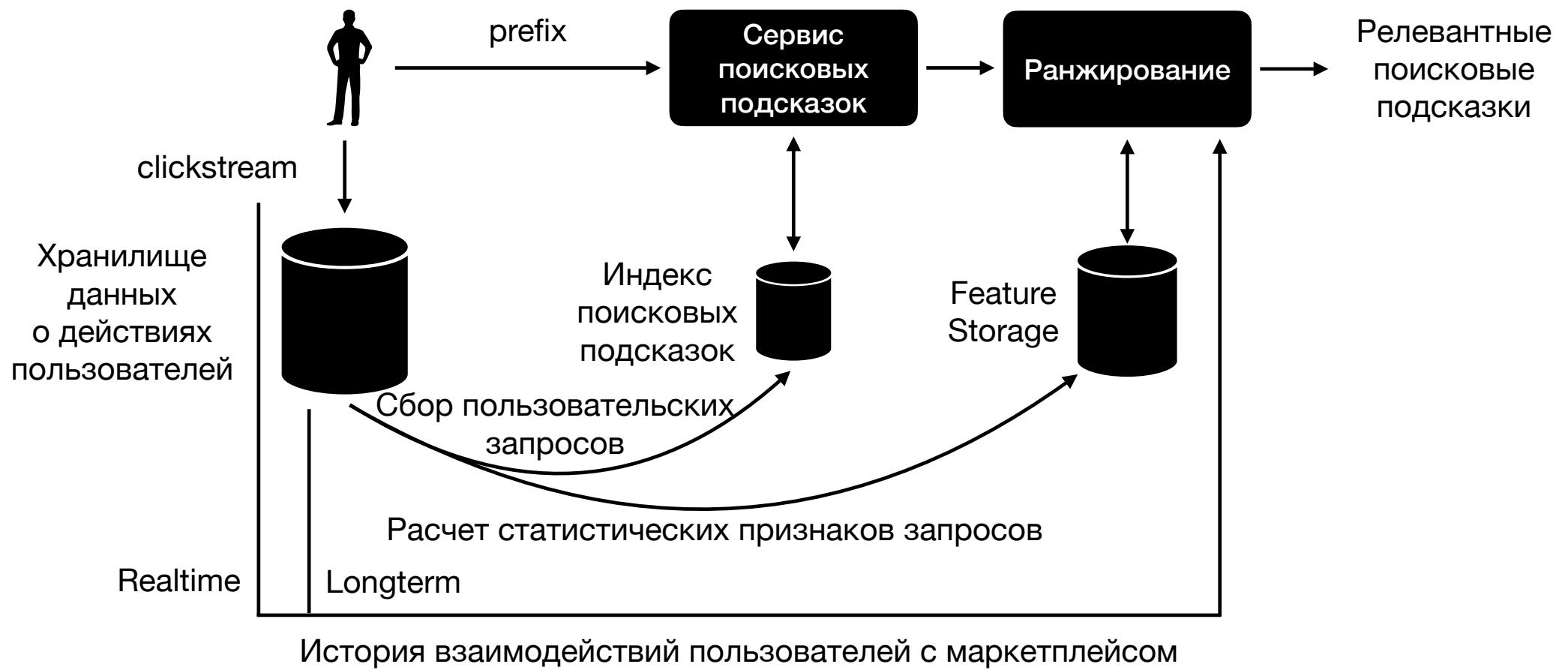
Признаковое описание подсказок

Нейросетевая персонализация

- Можно использовать те же идеи, что и в нейросетевых рекомендациях



Многоуровневая система подсказок



Что ещё важно учитывать?

Пользовательские признаки и предпочтения

- Регион пользователя
- Пол пользователя
- Ценовые предпочтения пользователя
- ...

Проблема опечаток

Проблема опечаток

- При вводе запроса (префикса) пользователи могут допускать опечатки

Проблема опечаток

- При вводе запроса (префикса) пользователи могут допускать опечатки
- **Проблема** – сейчас префиксный поиск этого никак не учитывает:
 - user prefix = «**кросов**»
 - search autocompletions = []

Проблема опечаток

- При вводе запроса (префикса) пользователи могут допускать опечатки
- **Проблема** – сейчас префиксный поиск этого никак не учитывает:
 - user prefix = «**кросов**»
 - search autocompletions = []
- Для решения этой проблемы можно:
 - обращаться в существующий сервис исправления опечаток (но префикс – недописанный запрос, может работать плохо!)
 - использовать нечеткий (fuzzy) поиск

Нечеткий поиск

- Обычно для оценки расстояния между строками используется расстояние Левенштейна
- Определяется как минимальное число операций (вставки символа, удаления символа, замены символа) необходимых для превращения одной строки в другую

Нечеткий поиск

Модификация Trie

- Будем обходить префиксное дерево с допущением опечаток
- Как это работает:
 - Поддерживается счетчик опечаток (например, максимум 2 ошибки, т.е. максимальное расстояние Левенштейна = 2)
 - На каждом шаге поиска допускается:
 - пропуск символа в префиксе (ошибка — удаление символа)
 - пропуск символа в дереве (ошибка — вставка символа)
 - замена символа на другой

Нечеткий поиск

Модификация Trie

- Пример:
 - prefix = «мало»
 - suggestions = [«**малоежка**», «**малосольные огурцы**», «**молоко**», ...]
- Проблемы:
 - вычислительная сложность поиска ()
 - проблема смешивания точных подсказок и fuzzy-подсказок (обычно решается на этапе переранжирования)

Генеративная постановка задачи

Недостатки текущего решения

- Мы ограничены конечной базой запросов — можем подсказать пользователю только те запросы, которые есть в нашей базе
- Исправление опечаток чаще всего решается исправлением опечаток / fuzzy-поиском, но и эти подходы не гарантируют успех
- Персонализация:
 - основывается в основном на handcrafted фичах
 - плохо учитываются долгосрочные предпочтения (нужно накрафтить очень много фичей)
 - разнородность пользовательского фидбека

Предпосылки генеративной постановки

- Пользователь вводит текстовый префикс
- Мы хотим предложить ему текстовую подсказку — последующие слова запроса

Предпосылки генеративной постановки

- Пользователь вводит текстовый префикс
- Мы хотим предложить ему текстовую подсказку — последующие слова запроса
- Очень похоже на задачу языкового моделирования:
 - оценить $P(w_i | w_1, w_2, \dots, w_{i-1})$, где
 - w_i — следующее слово (токен)
 - w_1, w_2, \dots, w_{i-1} — контекст (префикс)

Генеративная постановка задачи

- Пользователь вводит текстовый префикс
- Давайте учить **языковую модель**, которая будет генерировать продолжение префикса
- Будем решать недостатки текущего подхода:
 - можем показать подсказки практически по любому префиксу (не ограничены конечной базой подсказок)
 - исправлению опечаток можно научиться, собрав соответствующую обучающую выборку для модели

Генеративная постановка задачи

- Пользователь вводит текстовый префикс
- Давайте учить **языковую модель**, которая будет генерировать продолжение префикса
- Будем решать недостатки текущего подхода:
 - можем показать подсказки практически по любому префиксу (не ограничены конечной базой подсказок)
 - исправлению опечаток можно научиться, собрав соответствующую обучающую выборку для модели
- **Проблема:** продолжение префикса будет неперсонализированным

Генеративная постановка задачи

- **Проблема:** продолжение префикса будет неперсонализированным
- Модели слишком мало контекста в виде префикса, введенного пользователем
- **Предложение:** в качестве контекста можно использовать предыдущую историю поисков (запросы) пользователя

Генеративная постановка задачи

- **Проблема:** продолжение префикса будет неперсонализированным
- Модели слишком мало контекста в виде префикса, введенного пользователем
- **Предложение:** в качестве контекста можно использовать предыдущую историю поисков (запросы) пользователя
- N.B.: вообще говоря, можно использовать не только поисковую историю, но об этом потом

Генеративная постановка задачи

Search 1

Search 2

...

Search N

Пользовательская история запросов

Генеративная постановка задачи

Search 1

Search 2

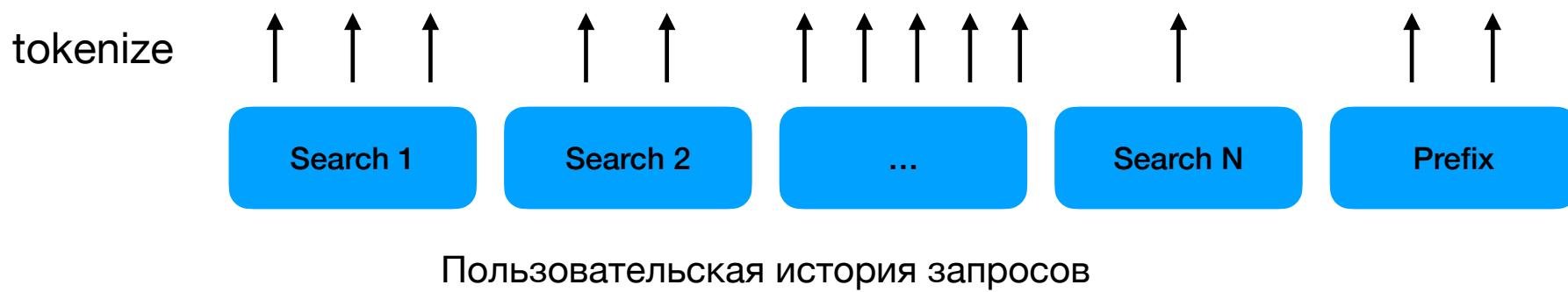
...

Search N

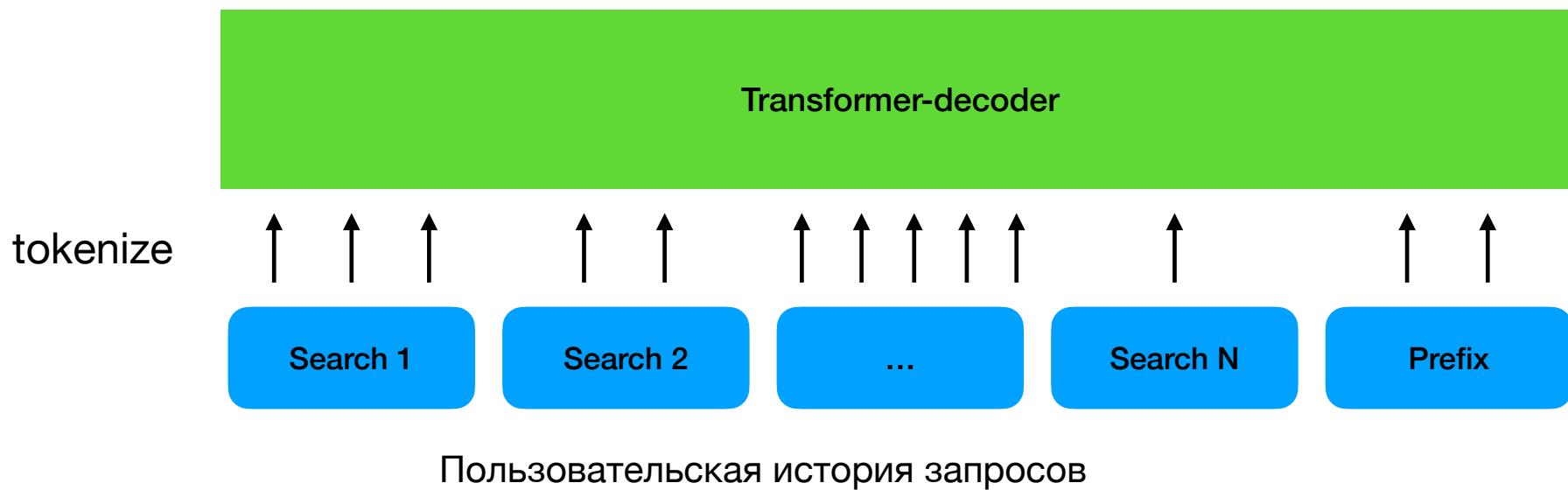
Prefix

Пользовательская история запросов

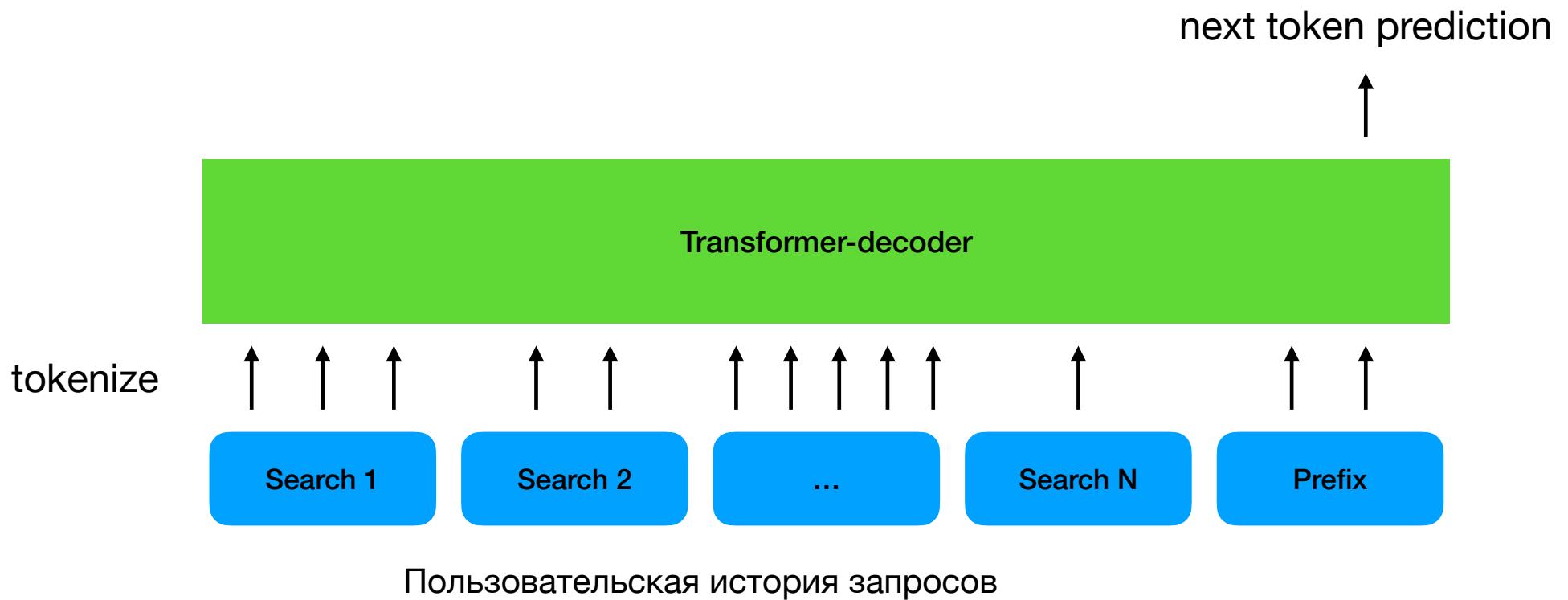
Генеративная постановка задачи



Генеративная постановка задачи



Генеративная постановка задачи



Генеративная постановка задачи

Преимущества

- **Решаем проблему покрытия:** Большое покрытие подсказками — практически для любого префикса мы сможем что-то сгенерировать и подсказать
- **Решаем проблему исправления опечаток:** Исправление опечаток работает «из коробки» в отличие от алгоритмического подхода (решается формированием выборки)
- **Решаем проблему персонализации** (ну почти): За счет механизма внимания подсказки основываются на предыдущей истории пользователя — персональные подсказки

Генеративная постановка задачи

Ограничения

- В силу жестких ограничений на время ответа модель должна инференситься очень эффективно:
 - либо не можем использовать модели большого размера
 - либо инференс в оффлайне с предрасчетом предсказаний модели (но тогда пропадает информация о текущем введенном префикссе)
 - либо комбинация двух этих подходов
- Учитываем только текстовую информацию из поисковых запросов
- Модель может генерировать:
 - то, чего мы найти не можем
 - обсценную лексику и экстремистские запросы

Эффективность генерации

- Использование модели относительно небольшого размера и дистилляция знаний большой модели в меньшую
- Фреймворки инференса (torch + cuda kernels, ONNX, TensorRT, TensorRT-LLM, vLLM, ...)
- Квантизация модели (bfloating16, int8, int4, ...)
- Использование KV-кеша
- Изменение архитектуры (GQA, MoE, MHLA)
- Стратегии декодинга (Speculative Decoding, Medusa, ...)
- ...

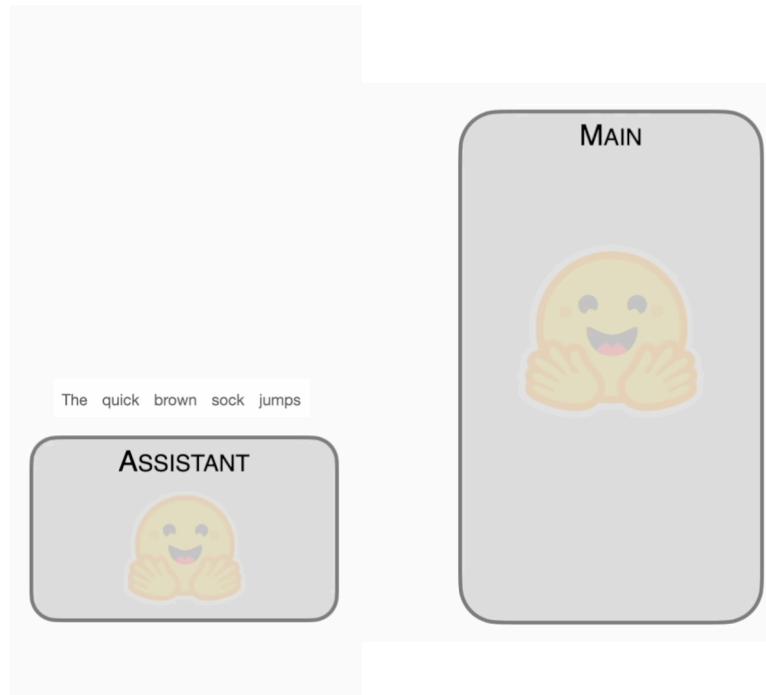
Декодинг

- Языковые модели by design умеют генерировать токены последовательно, один за другим:
 - Шаг 1: «[BOS]» -> «мама»
 - Шаг 2: «[BOS] мама» -> «мыла»
 - Шаг 3: «[BOS] мама мыла» -> «раму»
 - Шаг 4: «[BOS] мама мыла раму» -> «[EOS]»

Декодинг

Speculative decoding

- Появляется модель-ассистент (меньшего размера), которая генерирует продолжение, а большая (основная) модель верифицирует текст, который сгенерировал ассистент за **один** forward-проход



Декодинг

Speculative decoding

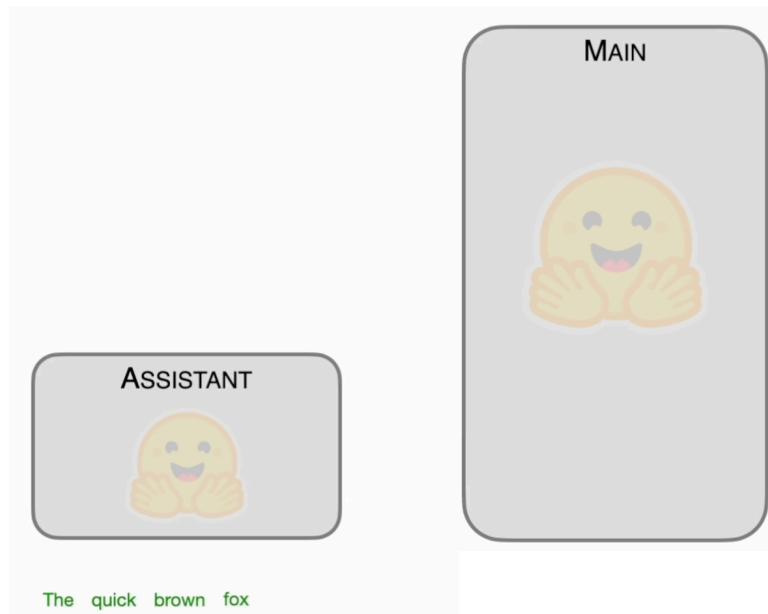
- Появляется модель-ассистент (меньшего размера), которая генерирует продолжение, а большая (основная) модель верифицирует текст, который сгенерировал ассистент за **один** forward-проход



Декодинг

Speculative decoding

- Появляется модель-ассистент (меньшего размера), которая генерирует продолжение, а большая (основная) модель верифицирует текст, который сгенерировал ассистент за **один** forward-проход



Декодинг

Speculative decoding

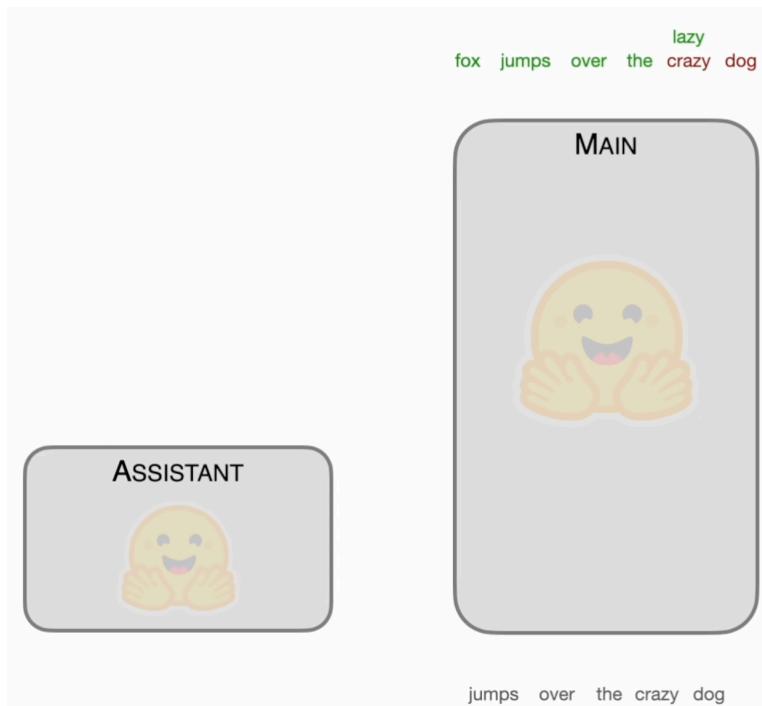
- Появляется модель-ассистент (меньшего размера), которая генерирует продолжение, а большая (основная) модель верифицирует текст, который сгенерировал ассистент за **один** forward-проход



Декодинг

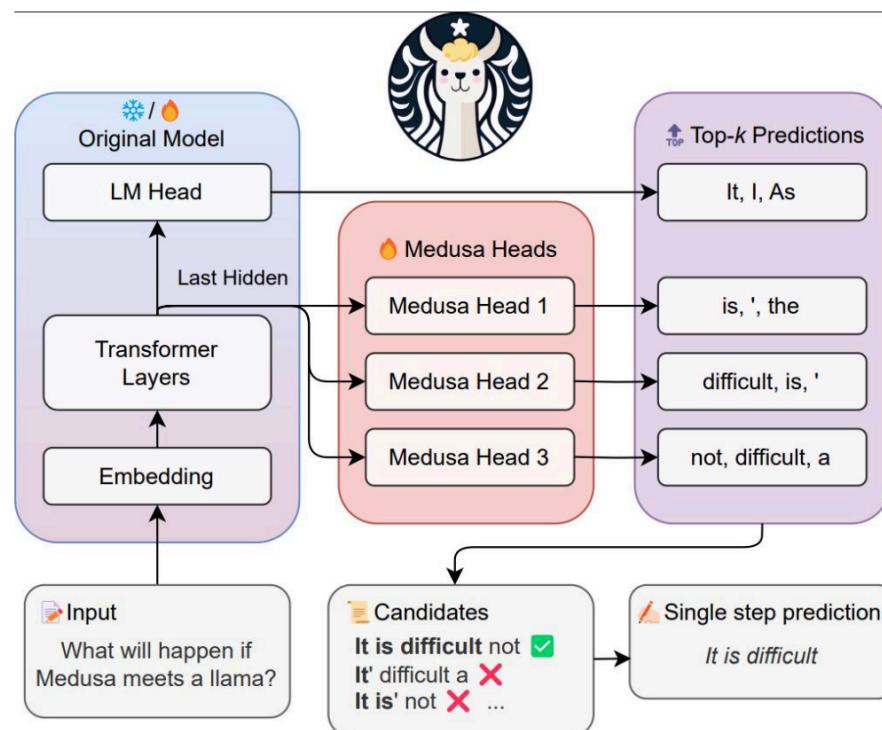
Speculative decoding

- Появляется модель-ассистент (меньшего размера), которая генерирует продолжение, а большая (основная) модель верифицирует текст, который сгенерировал ассистент за **один** forward-проход



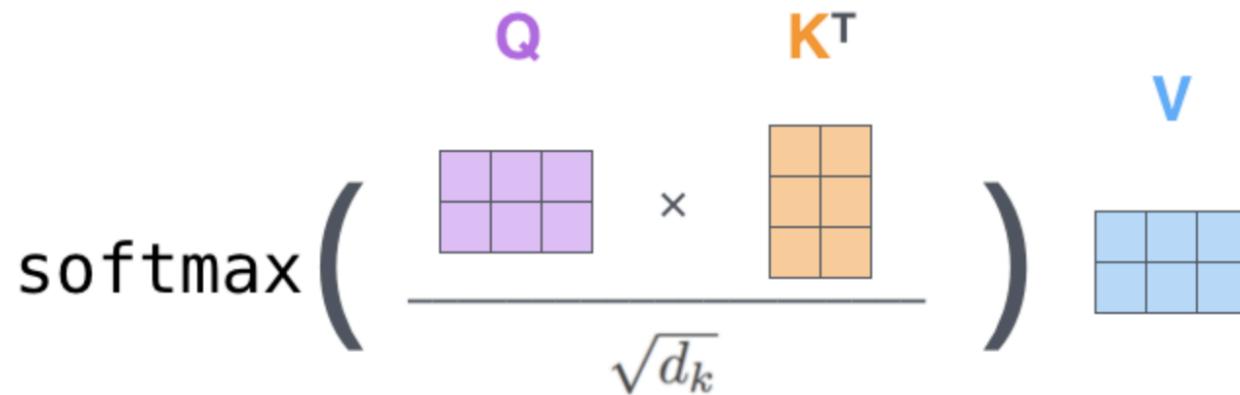
Декодинг Medusa

- Для уже обученной LLM присоединяем дополнительные головы, которые будут генерировать дальнейшее продолжение



Эффективность генерации KV-кеш

- Идея в том, что можно закешировать уже посчитанные матрицы K и V

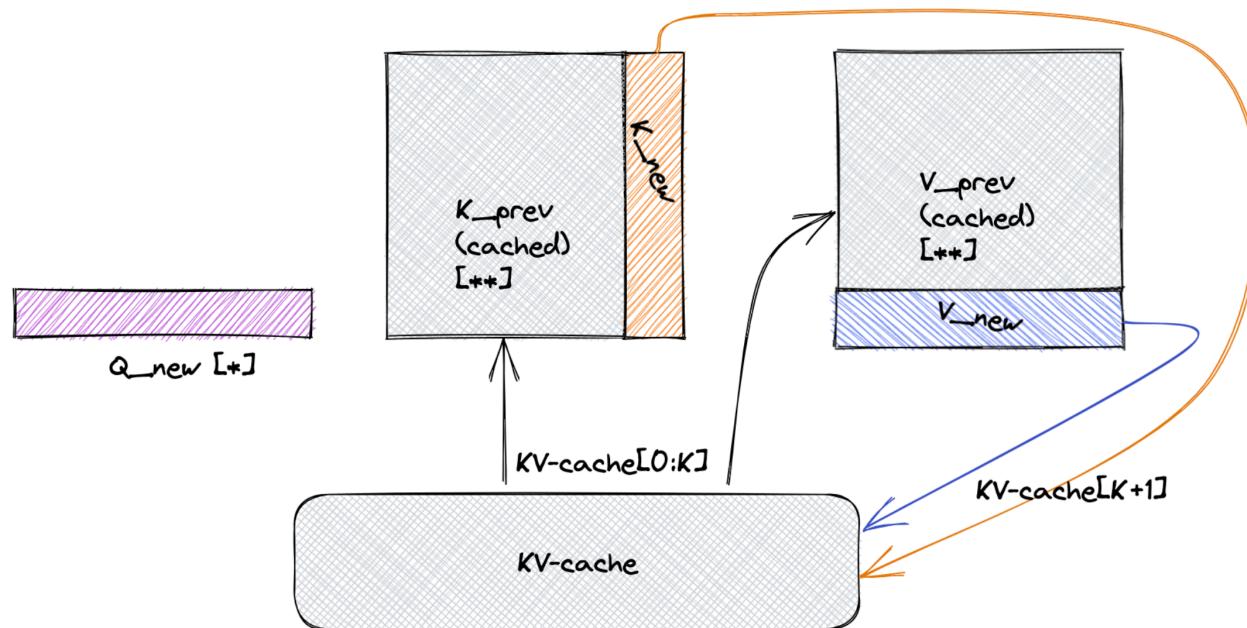
$$\text{softmax} \left(\frac{\begin{array}{|c|c|c|} \hline & Q & K^T \\ \hline & \times & \\ \hline \end{array}}{\sqrt{d_k}} \right) V$$


<https://huggingface.co/blog/not-lain/kv-caching>

<https://medium.com/@joaolages/kv-caching-explained-276520203249>

Эффективность генерации KV-кеш

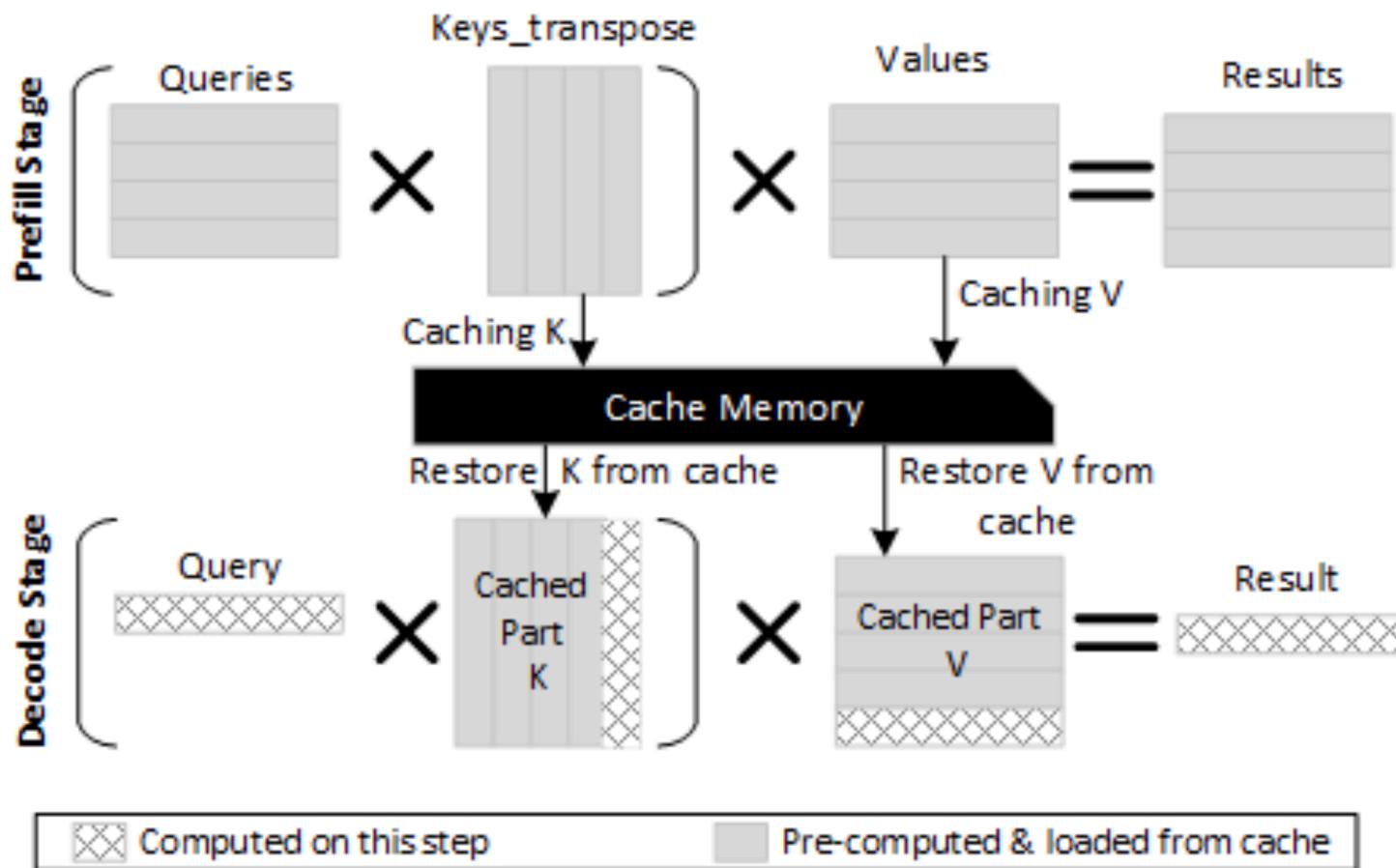
- Идея: можно закешировать уже посчитанные матрицы K и V



<https://huggingface.co/blog/not-lain/kv-caching>

<https://medium.com/@joaolages/kv-caching-explained-276520203249>

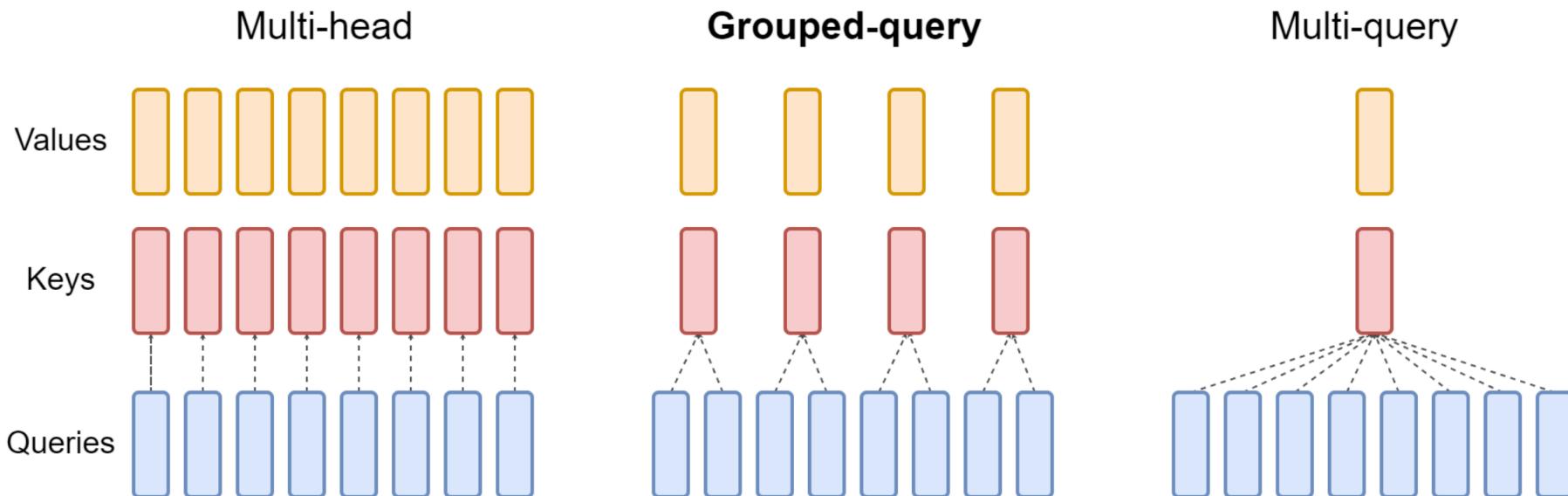
Эффективность генерации KV-кеш



Эффективность генерации

Изменение архитектуры – Grouped Query Attention (GQA)

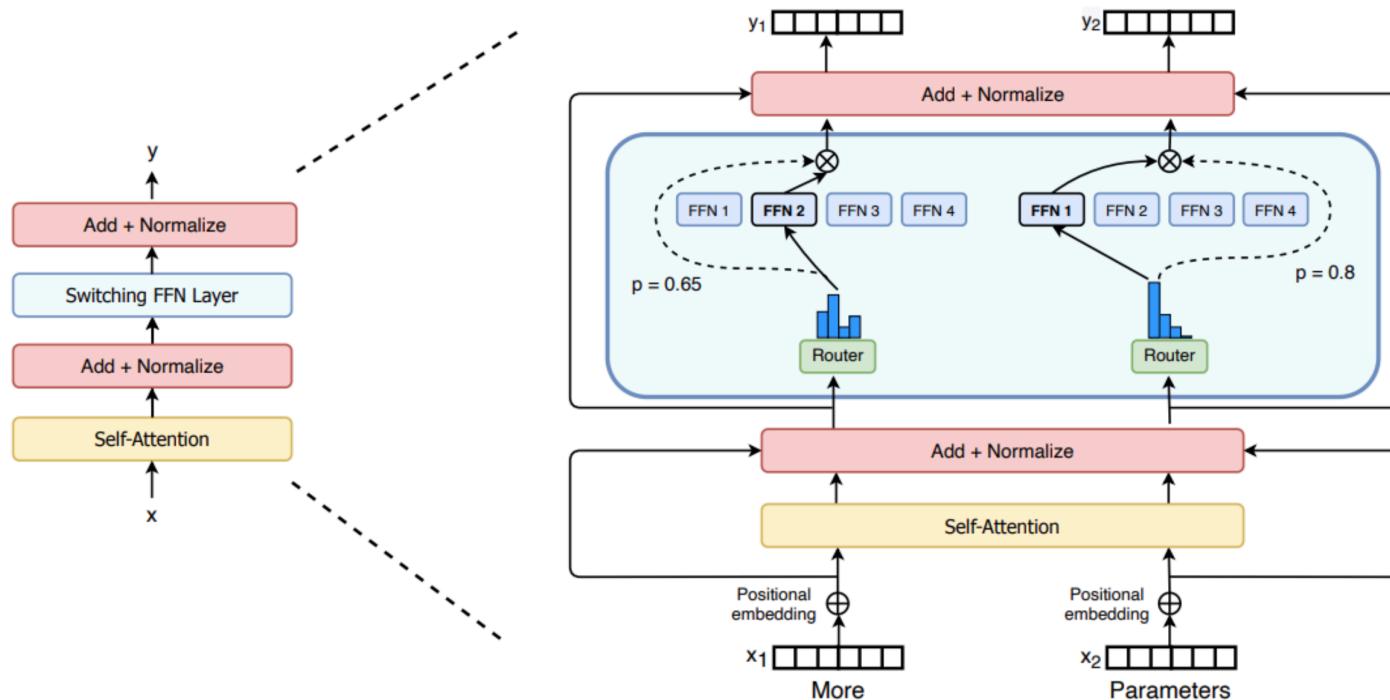
- Идея: нескольким Q соответствуют одни и те же K и V



Эффективность генерации

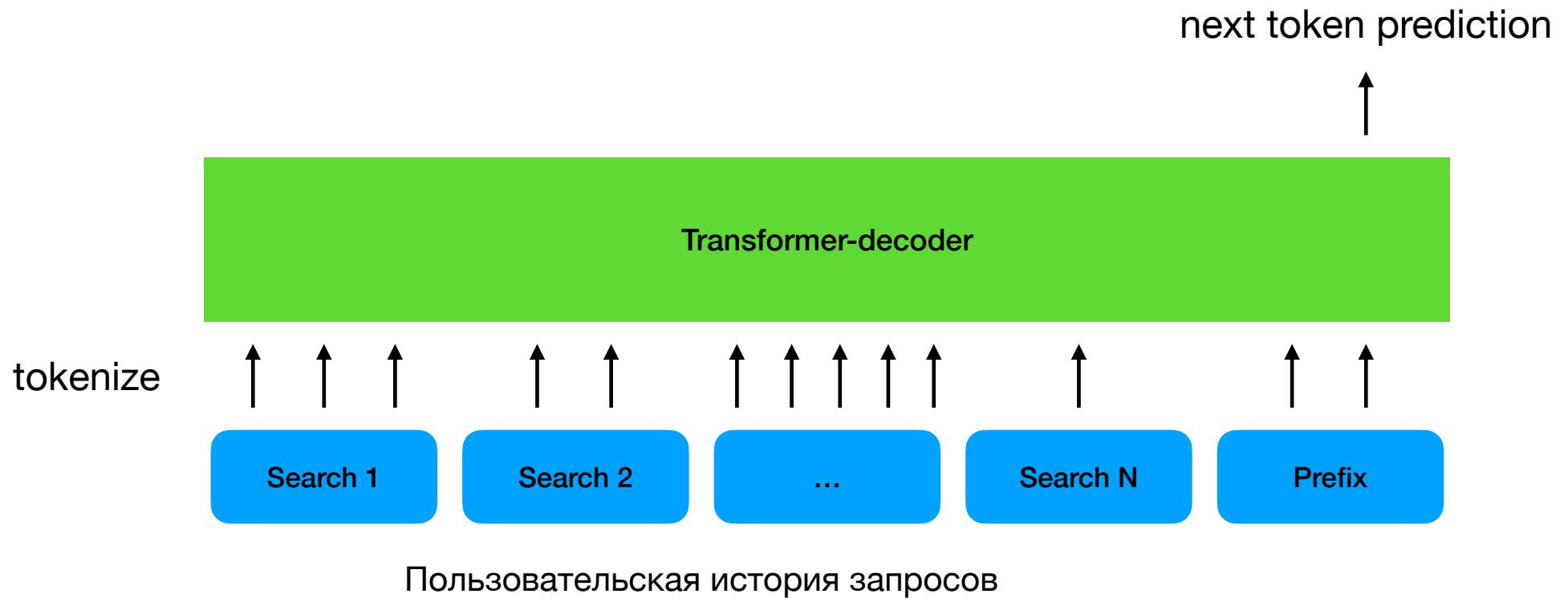
Изменение архитектуры – Mixture of Experts (MoE)

- Идея: искусственно увеличиваем число параметров модели при том же числе «активных» параметров

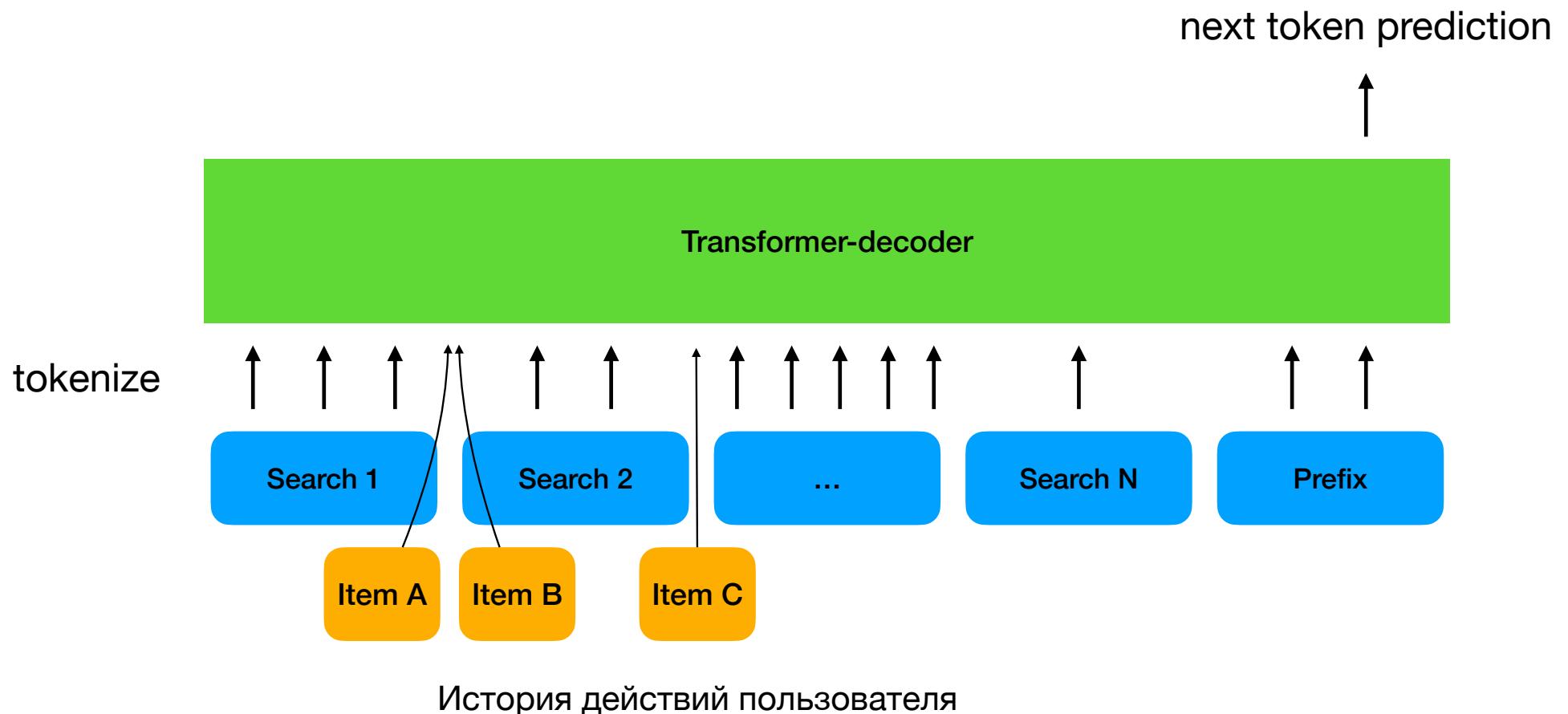


<https://huggingface.co/blog/moe>

Учет нетекстовой информации

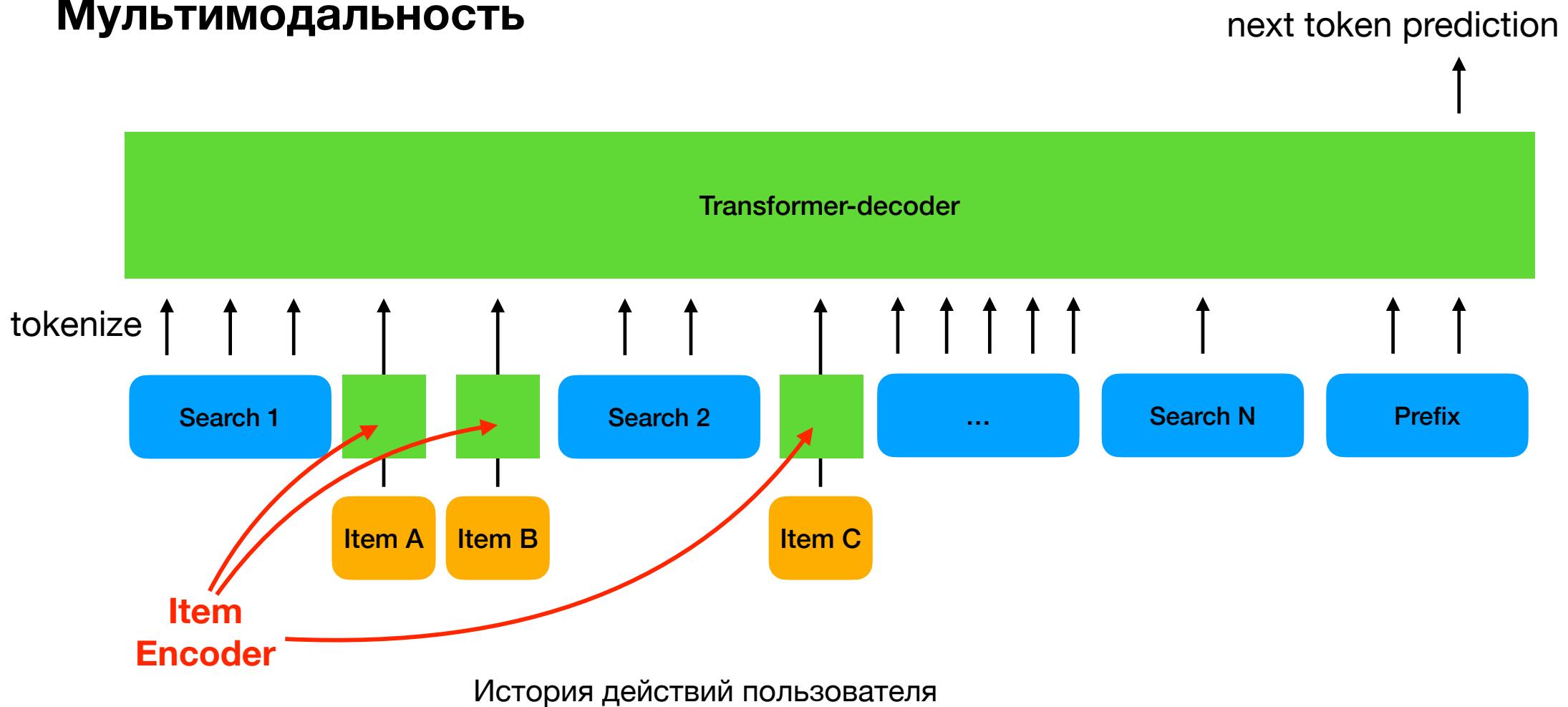


Учет нетекстовой информации



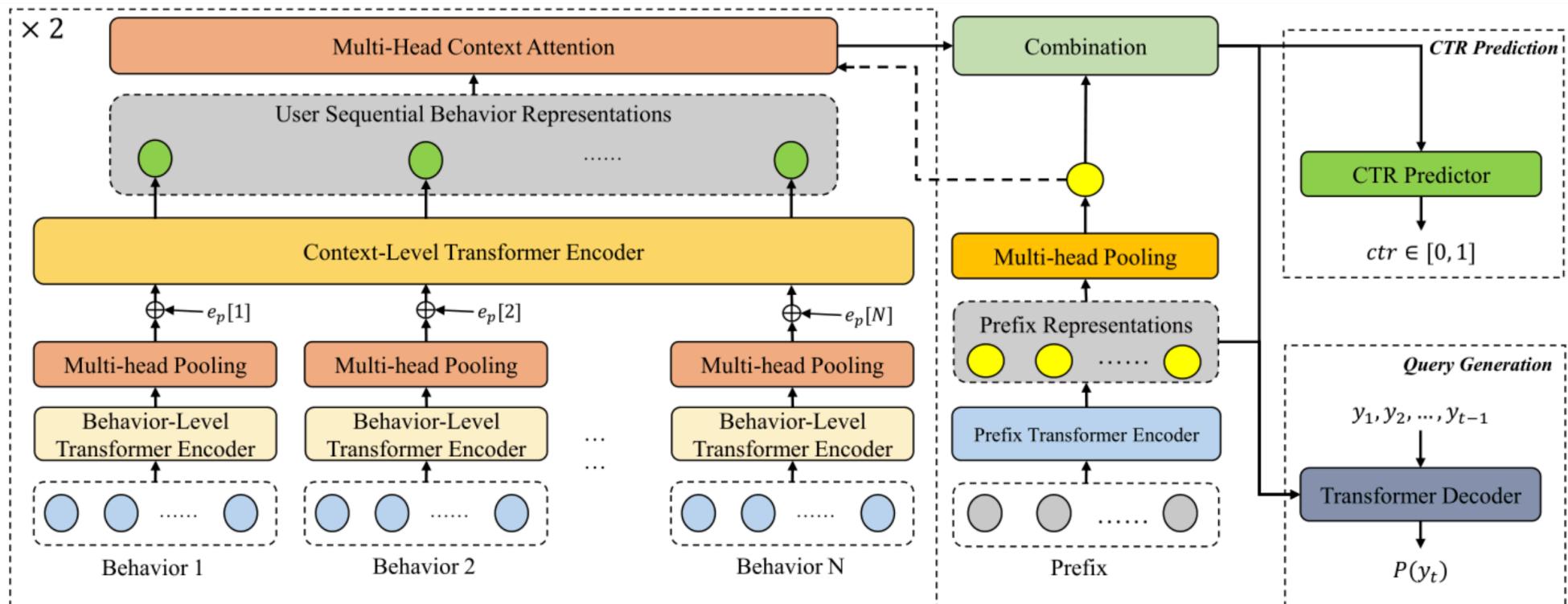
Учет нетекстовой информации

Мультимодальность



Учет нетекстовой информации

Encoder-decoder



https://www.alibabacloud.com/blog/kdd-2020-learning-to-generate-personalized-query-auto-completions-with-a-new-approach_596752

Решение прочих проблем

- Генерация того, чего мы не сможем найти, то есть ввод запроса не даст результата поиска:
 - можно проверять перед показом подсказки пользователю, а найдем ли мы что-то
 - можно решать проблему балансировкой выборки — учиться на конверсионных запросах
- Генерация обсценной и экстремистской лексики:
 - бан по словарю / правилам (но словарь нужно иметь)
 - стадия alignment, когда мы говорим модели в лоссе, как надо, а как не надо генерировать

Другие downstream-задачи

- Генерация не только текстовых подсказок
 - категории
 - фильтры
- Исправление опечаток
- Рекомендации запросов
- Использование как базовой модели пользователя для персонализации в различных сценариях

Поисковые подсказки

Андронов Дмитрий, 19.05.2025, AI Masters