Яндекс

Лингвистика длинного хвоста

Николай Григорьев

Отдел голосовых технологий

Устройство Web-поиска

Индекс:

- архив документов
- обратный индекс: по слову выдает все содержащие его документы
- данные о документах
- данные о сайтах

Устройство Web-поиска

- Разбор запроса:
  - ключи для поиска в обратном индексе
  - факторы для ранжирования
- Фильтрация:
  - поиск в индексе по ключам
- Ранжирование:
  - выбор наилучшего результата
- Извлечение сниппетов
- Отрисовка страницы результата

Морфология

- Задача – извлекать из индекса документы по всем формам слова из запроса
- Ключи для поиска:
  - хранить в индексе словоформы:
    - по словоформе из запроса определять остальные;
  - хранить в индексе леммы:
    - по словоформе из запроса определять лемму.
- Частотная лексика: словарь Зализняка

Морфология

- Зализняк: ~100 тыс. входов, ~3-4 млн словоформ, 85-90% текста
  - Поисковый индекс: ~200 млн словоформ – **длинный хвост**
    - Как обеспечить морфологией слова не из словаря?

Морфология

«Глокая куздра штеко будланула бокра и кудрячит бокренка.» (Л.В.Щерба)

глокая – глокий? глокать? глокай?

куздра – куздра? куздр?

штеко – штекий? штеко?

бокра – бокр? бокра? бокрый?

кудрячит – кудрячит? кудрячитый? кудрячить?

бокренка – бокренк? бокрёнок? бокренка?

Морфология

Как найти лучший разбор?

- По самому длинному совпадению со словарным словом (Илья Сегалович)
- Машинное обучение:
  - Частоты отдельных окончаний и основ, вероятность данного окончания при основе
  - Словарь частотных морфологических единиц в качестве обучающего множества
- Снятие неоднозначности по контексту
- Автоматическая морфология ☺.

Исправление опечаток

- 10-15% запросов содержат опечатки
- Разные типы опечаток
  - ошибки набора
  - орфографические ошибки
  - неправильная раскладка (*lytdybr*)
- Самые частотные (*агенство*) можно занести в словарь, остальные надо предсказывать – опять длинный хвост!

Исправление опечаток

- Ищем **гипотезы** исправления
- Взвешиваем два фактора:
  - насколько трудно было так ошибиться: **функция ошибки**
  - насколько хороший текст получается при исправлении: вероятность гипотезы по **языковой модели**.

Исправление опечаток

- Начало:
  - фиксированный словарь хороших слов
  - вручную подобранная функция ошибки на базе расстояния Левенштейна (edit distance);
  - триграммные языковые модели
- Современный уровень:
  - словарь гипотез автоматически извлекается из корпусов документов и запросов
  - функция ошибки подбирается методами машинного обучения
  - продвинутые языковые модели – большой размерности, нейронные сети и т.п.

Расширение запроса

Найти другие способы выразить тот же смысл

словообразование

[москва метро] ↔ [московское метро]

аббревиатуры

[рф] ↔ [российская федерация]

транслитерация

[кофеварка бош] ↔ [кофеварка bosch], [почта яндекс] ↔ [yandex mail]

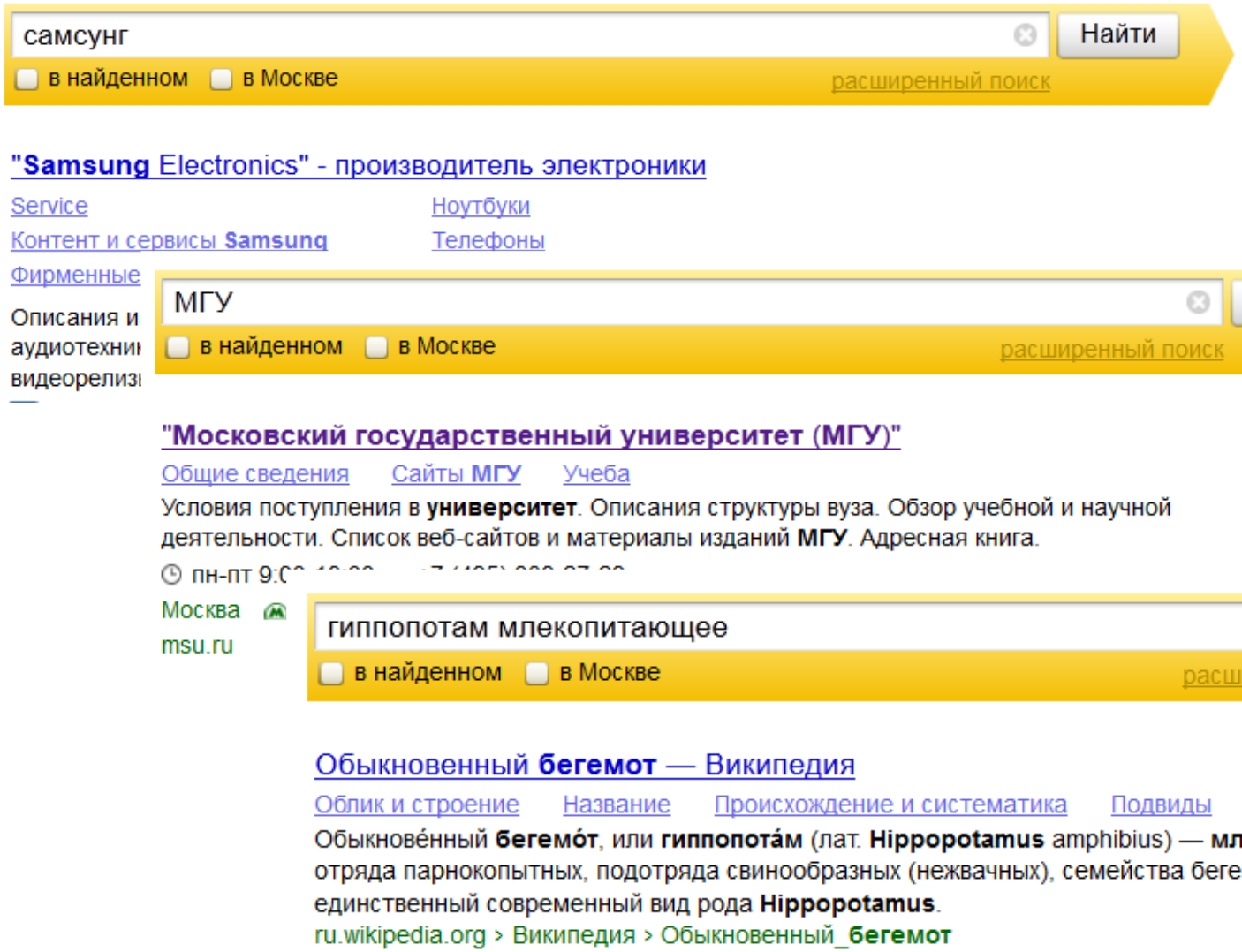
орфографические варианты

[икеа химки] ↔ [икея химки], [смотреть онлайн] ↔ [смотреть он-лайн]

синонимы

[гиппопотам фото] ↔ [бегемот фото]

Как заметить расширения?



Расширение запроса

- Проблема длинного хвоста острее, чем в морфологии: от существующих словарей еще меньше проку
- Расширения **зависят от контекста**; описать эту зависимость правилами – нереально.

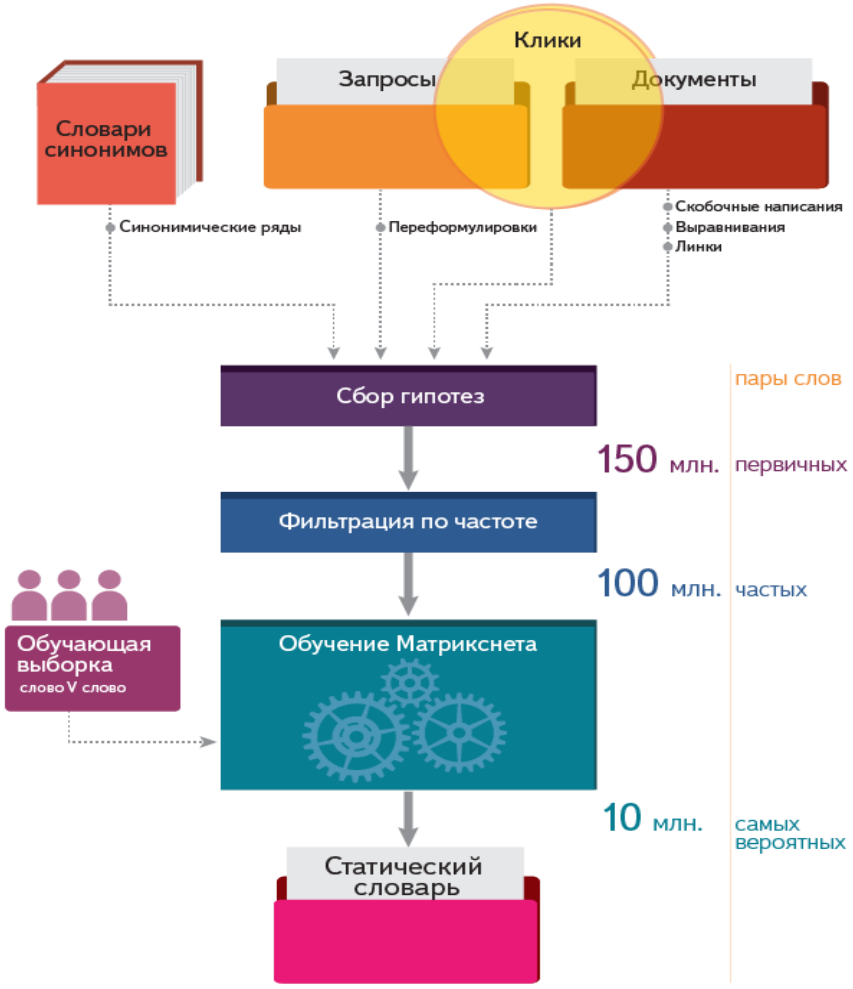
Контекст запроса: когда нельзя расширять

Расширение запроса

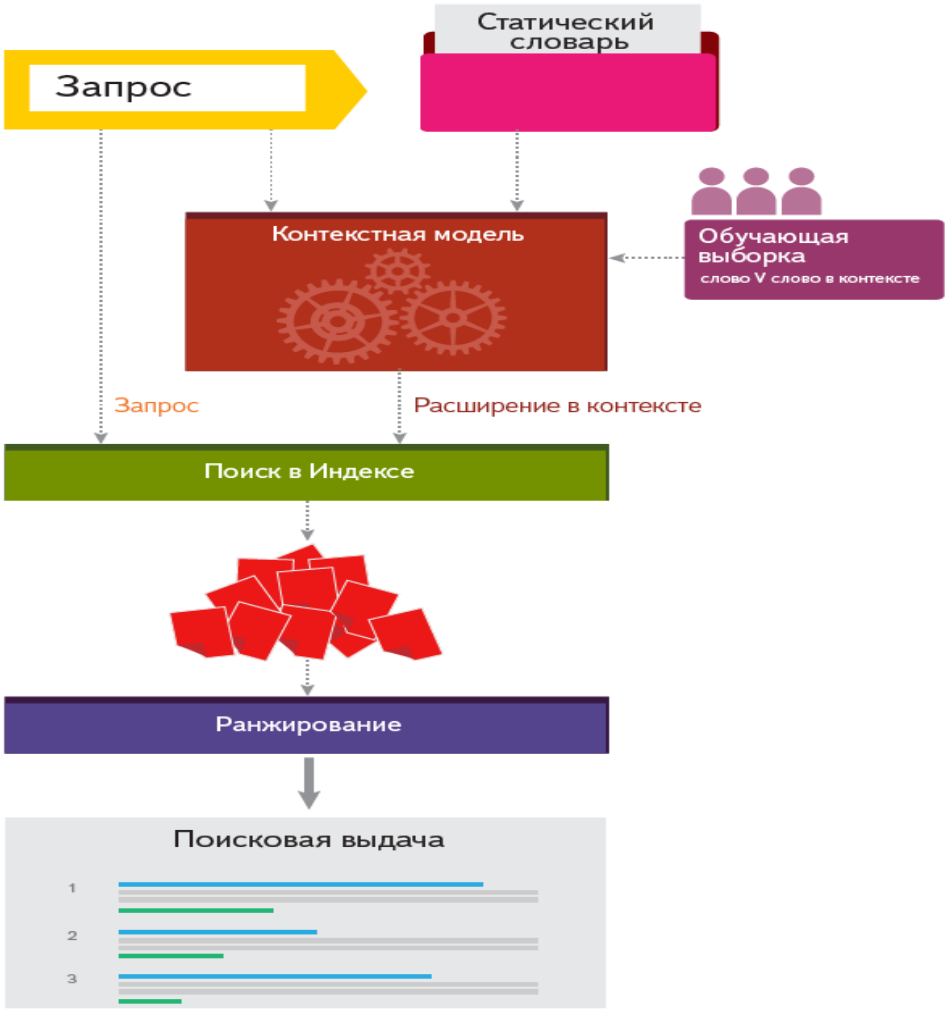
Выход – машинное обучение:

- сбор гипотез расширения – редиректы Википедии, скобочные написания, переформулировки запросов, данные о пользовательских кликах, …
- составление эталонной разметки
- генерация факторов
- обучение классификатора

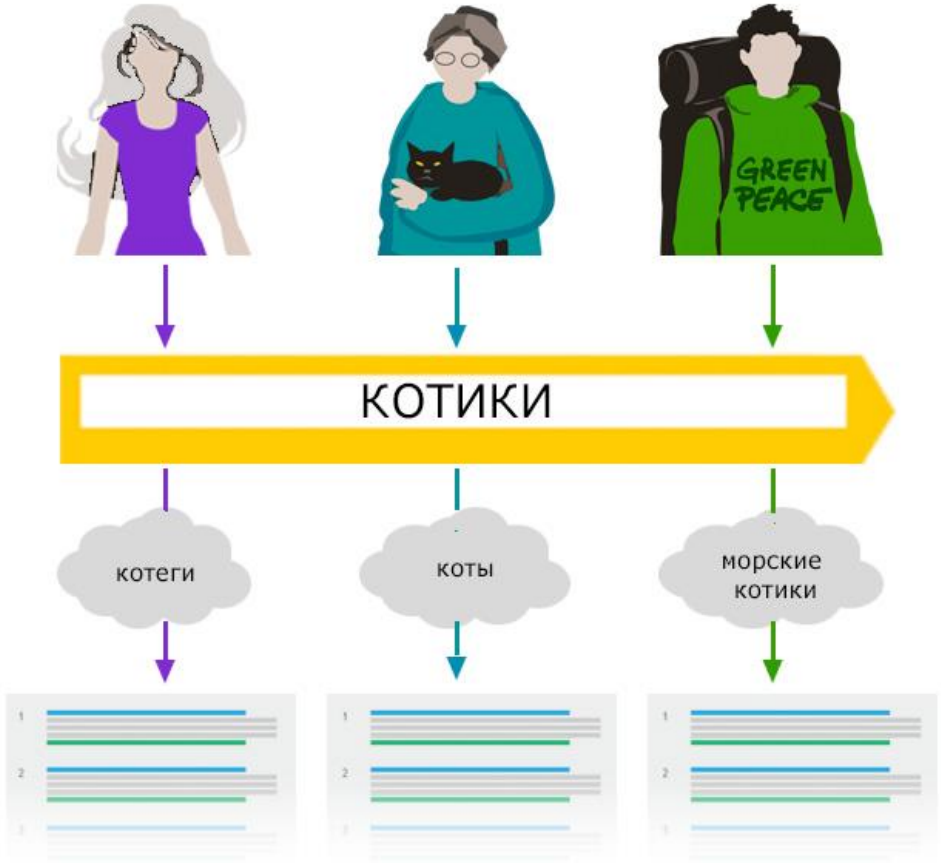
Подготовка данных для расширений



Расширение в контексте запроса



Персонализируем расширения



Определение языка

- Язык документа: определяет морфологию, влияет на релевантность, позволяет группировать и фильтровать документы

- Язык запроса: определяет морфологию, опечатки, расширения и вообще все.

Почему трудно понять язык запроса?

Запросы короткие: в среднем около 3 слов

Лексика и синтаксис отличаются от нормативной речи

Опечатки: 10-15%

Неоднозначная интерпретация:

[**ISO-8859-9**] для запроса из России: RU или EN
[**Facebook**] для запроса из Турции: скорее TR, чем EN
[**Приват-банк**] для запроса из Украины: RU или UK

Язык запроса не совпадает с языком искомого документа:

[**El condor pasa**] — английская песня Simon&Garfunkel

Смесь языков:

[**pretty little liars 3. sezon kaç bölüm olacak**]
[**Скачать сочинение на тему: Чи настане той час, коли в кожній українській родині спілкуватимуться рідною мовою**]

Распознавание языка запроса

Информация о тексте

письменность (кириллическая, латиническая, арабская); набор символов, специфичных для конкретного языка; характерные слова языка; буквенные n-граммы

Информация о пользователе:

местонахождение (по IP-адресу); язык поискового интерфейса; домен верхнего уровня (yandex.ru, yandex.ua, yandex.com.tr)

Машинно-обученный классификатор — каждый язык определяется своей комбинацией факторов

[**дружина князя игоря**] — ищем информацию о войске,
[**дружина князя ігоря**] — ищем княгиню Ольгу и Ярославну

А также

- Выделение имен и географических названий
- Выделение объектов в запросах и документах
- Тематическая классификация документов и запросов

Всюду длинный хвост

- Лингвистика в Яндексе – это не описание известных вам языковых единиц, а построение систем, могущих справиться с неизвестными
- Есть очень много неразмеченных данных и вычислительных ресурсов для работы с ними.
- Размеченных данных мало и они дороги.
- Доминирующие подходы – машинное обучение и вероятностное языковое моделирование.
- Необходимые умения – анализ больших данных, математическая статистика и вероятностные методы; в ближайшем будущем – нейронные сети.

Яндекс

Николай Григорьев

grig@yandex-team.ru

Спасибо!