Теория информационного поиска

Лекция 7. Odds and ends

Дмитрий Грановский

СПбГУ

05.05.2025

Расширение запросов

Решаем проблему:

- в очевидно релевантном документе термин не совпадает с термином из запроса
- и разница не сводится к нормализации
  - [голландия достопримечательности] vs “достопримечательности Нидерландов”
  - [список альбомов металлики] vs “Metallica, все альбомы”

Расширение запросов

Общая идея:

- заведем словарь терминов, у которых есть регулярные «синонимы»
- если находим такой термин в запросе, то приписываем «синоним» через ИЛИ:
  - `[(голландия ^ нидерланды) достопримечательности]`
- это называется *расширением запросов* (query expansion)
- структура запроса — дерево

Типы расширений

- ~синонимы: страна/государство, ремонтировать/починить
- словообразовательные: физика/физический
- транслиты: металлика/metallica, бош/bosch
- аббревиатуры: ип/индивидуальный предприниматель
- склейка/разрезание: авто кредит/автокредит (см. орфоварианты)

Расширения: проблемы

- неоднозначность
- асимметрия
- пример: [bosch] -> {босх | бош}
  - снимается контекстом:
    - [hieronymus bosch] -> босх
    - [шуроповерт bosch] -> бош
  - [бош] -> [bosch] ок,
    - [босх] -> [bosch] ок?
- контекст может быть нетекстовым: [МГУ] зависит от региона

Расширения: сбор

Общий подход:

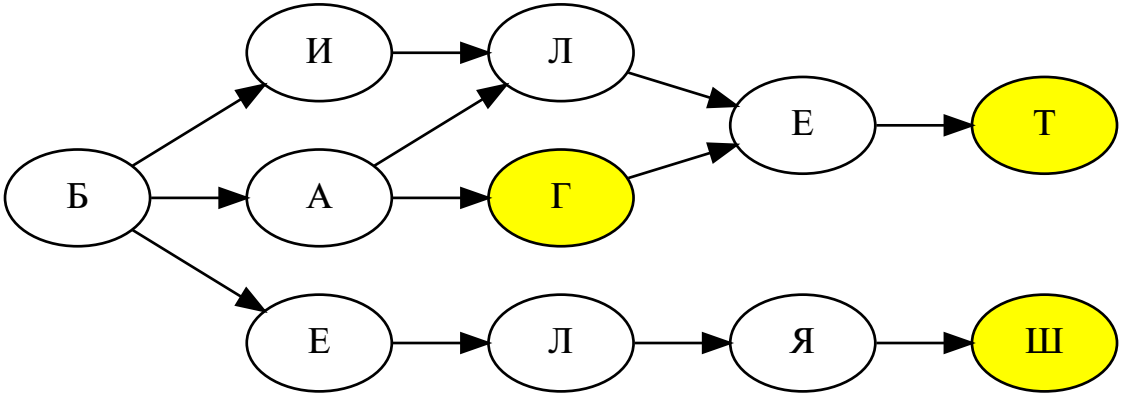
- исходные данные: корпусы, история запросов, история переформулировок
- генерация гипотез
- фильтрация
  - по лингвистической модели (тезаурус, модель словообразования, правила транслитерации, …)
  - по статистике (т.к. хорошая пара не всегда хорошее расширение, напр. [магазин]/[magazine])
- доочистка вручную или ML-моделью

Поиск по маске (wildcard)

Запросы вида: `[анти*изм]`, `[рос*надзор]`

- в веб-поиске — слишком (вычислительно) дорого
- иногда расширяется до поиска по регулярным выражениям
- потенциальные применения:
  - сомнение в орфографии: `[вин*грет]`
  - неполная информация: `[к*ский история россии]`
  - поиск >1 объекта: `[сульф* меди]`
  - …

Trie (префиксное дерево, бор)



Поиск по маске, подход 1

- при индексации строим дерево поиска, например, trie
  - так можем искать по любому префиксу: `[сульф*]`
- также можно построить обратное дерево
  - теперь можно искать по постфиксу: `[*изм]`
  - и с двух сторон: `[вин*грет]`
- результат поиска по дереву/деревьям затем ищем в обычном индексе

Поиск по маске, подход 2

- при индексации разбиваем термин на N-граммы
- винегрет = ^ви + вин + ине + нег + егр + гре + рет + ет$
- строим отдельный N-граммный индекс, ключу соответствует список терминов
- `[вин*грет]` = `^ви AND вин AND гре AND рет AND ет$`
  - возможно, с дополнительной фильтрацией найденных гипотез

Сниппеты

- также известны как *поисковые аннотации*
- часть документа, показываемая на странице выдачи (SERP)
- *динамические* сниппеты: зависят от запроса
- плюсы:
  - дают представление о документе до клика
  - иногда решают всю задачу (особенно для фактовых запросов)
- минусы:
  - занимают ценное место на экране
  - требуют времени на генерацию
  - плохо понятно, как измерять качество

Разнообразие и свежесть

В остальных лекциях качество выдачи $\approx$ релевантности

- однако некоторые свойства хорошей выдачи трудно сформулировать как просто оценки релевантности
- свежесть: актуальные события, меняющиеся параметры
- разнообразие: когда не уверены в *интенте*
  - [властелин колец] — книга или фильм?
  - [сталкер] — фильм? игра? группа?
  - но: осторожно с adult интентами
- мультимедийность: нужно ли показать картинки/видео
  - но: некоторые изображения не стоит показывать без предупреждения