Теория информационного поиска

Лекция 1. Введение

Дмитрий Грановский

СПбГУ

10.03.2025

Устройство курса

- слайды выдаются
- 1 семестр, в конце зачет
- 10 лекций + 4 практики
- оценка из 100 баллов
- для зачета надо 60 баллов
- где взять баллы:
  - 70 баллов — д/з
  - 30 баллов — зачет
  - 0 баллов — посещение
- д/з — тесты и программирование (Python)
  - у них есть дедлайны

Ваша обратная связь важна
и приветствуется!

Почитать

1. Маннинг, Рагхаван, Шютце. Введение в информационный поиск.
2. Лукашевич. Тезаурусы в задачах информационного поиска.
3. Грановский. Информационный поиск // Прикладная и компьютерная лингвистика (ред. Ландо и др.)
4. Логинов. Задача ранжирования // Яндекс. Учебник по машинному обучению: https://clck.ru/3Absui
5. Bruce Croft, Metzler, Strohman. Search Engines: Information Retrieval in Practice.

Посмотреть

1. Ян Кисель и др. Курс лекций от Mail.ru, 2017. https://www.youtube.com/playlist?list=PLrCZzMib1e9o_BlrSB5bFkLq8h2i4pQjz
2. Илья Марков. Курс лекций (CSCenter), 2016. https://www.youtube.com/playlist?list=PLlb7e2G7aSpQJH3UefgoBBxkCz0xn2R3x
3. Алексей Зобнин. Лингвистика в поиске (Яндекс), 2013. https://youtu.be/zGjzUGOMKgc
4. Петр Попов и др. Как устроен поиск Яндекса: о чем невозможно прочитать, 2016. https://youtu.be/BCVsgup8hUQ

Содержание курса

1. Введение (сегодня)
2. Классические модели поиска (2, 4)
3. Ранжирование (7, 9)
4. NLP (6, 10, 13)
5. Оценка качества поиска (11)
6. Архитектура веб-поиска (14)
7. +Пишем поисковую систему на Python (3, 5, 8, 12)

Что такое поиск?

- билеты по дате + направлению?
- книга в библиотеке по автору?
- фильм в онлайн-сервисе по году и жанру?

Что такое поиск?

- билеты по дате + направлению?
- книга в библиотеке по автору?
- фильм в онлайн-сервисе по году и жанру?

Всё это поиск, но не наша тема.

Разные данные

1. Структурированные данные
   - например, база данных или каталог
   - запрос специального вида (например, SQL)
   - все примеры выше были этого типа
2. Неструктурированные данные
   - чаще всего текстовые
   - самое привычное нам — интернет (веб-поиск)
   - обычные текстовые запросы (ad hoc search)
     - тж. *полнотекстовый поиск*
3. Полуструктурированные

Пример SQL-запроса

```sql
SELECT
    title, director
FROM
    films
WHERE
    genre IN ('боевик', 'экшн', 'драма')
    AND year > 2004
    AND country != 'Китай'
    AND score >= 7.0
ORDER BY score DESC
```

Терминология

1. Информационный поиск
   - Information Retrieval
2. Поисковая система
   - Search Engine
3. Поисковый запрос
   - Search Query

Где ищем?

- информация находится в **документах**
  - обычно текстовых:
    - страница в интернете,
    - документ на компьютере,
    - сообщение в мессенджере, …
  - но в принципе документом может быть:
    - картинка,
    - видеоролик,
    - фрагмент речи, …
- все документы в совокупности — **коллекция**

Что ищем?

- **Информационная потребность** (*information need*) — то, что пользователь хочет найти.
  - похожее понятие — **интент** (*intent*)
- **Запрос** (*query*) — то, как он выражает эту потребность.

Что ищем?

- **Информационная потребность** (*information need*) — то, что пользователь хочет найти.
  - похожее понятие — **интент** (*intent*)
- **Запрос** (*query*) — то, как он выражает эту потребность.

Как и документ, запрос может быть картинкой и т. д.

Типы запросов (и инф. потребностей)

1. информационные:
   - фактовые:
     - [высота джомолунгмы]
     - [какого числа пасха в 2024]
     - [бжу пельмени]
   - широкие
     - [заполнение декларации]
2. навигационные
   - [сайт филологического факультета спбгу]
3. транзакционные
   - [купить билет питер хабаровск ржд]

Из чего состоит документ?

- из терминов/термов (*terms*)
- чаще всего это слова, но некоторые строки трудно назвать словами:
  - 10.03.2025
  - mail@example.com
  - Боинг-747
  - C++
  - [вставьте название своей музыкальной группы]

Релевантность

1. Субъективное понятие, описывающее степень «подходящести» ответа (в т. ч. документа) к информационной потребности
   - это **пользовательская релевантность**
2. Численное выражение того, насколько поисковая система считает ответ (документ) подходящим к запросу
   - это **вычисленная релевантность**

Релевантность

1. Субъективное понятие, описывающее степень «подходящести» ответа (в т. ч. документа) к информационной потребности
   - это **пользовательская релевантность**
2. Численное выражение того, насколько поисковая система считает ответ (документ) подходящим к запросу
   - это **вычисленная релевантность**

Определение того, что такое релевантность и как её считать — одна из центральных теоретических проблем IR.

Резюме

1. Рассматриваем поиск по текстовой *коллекции*.
2. Коллекция состоит из *документов*.
3. Документ состоит из *терм(ин)ов*.
4. Пользователь имеет *информационную потребность* и задает *запросы*.
5. То, насколько документ подходит к запросу, — *релевантность*.