МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)
Факультет информационных технологий и прикладной математики
Кафедра вычислительной математики и программирования

Отчет по лабораторной работе №2
«Оценка качества поиска»
по курсу
«Информационный поиск»

Группа: 80-106М
Выполнил: Демин И.А.
Преподаватель: Калинин А.Л.

Москва, 2019

Задание

Необходимо оценить качество своего поиска и сравнить их с двумя альтернативами (для Википедии можно собственный поиск по Википедии, поиск Google или Яндекса с ограничением по сайту Википедии). Как минимум, нужно измерить P, DCG, NDCG и ERR уровней @1, @3 и @5, приветствуется использование дополнительных метрик качества.

Для оценки качества необходимо придумать 30 запросов, отражающих интересы пользователей или, если есть доступ к настоящим запросам пользователей, то выбрать репрезентативную подборку. В качестве примера посмотрите на 10 запросов к поиску по всей Википедии, подумайте о том, почему именно они были выбраны и какую сложность для поисковой системы они представляют:

Выполнение лабораторной работы

Для сравнения были взяты поисковые системы Google с ограничением по сайту wikipedia.org и поиск самой Wikipedia. Поисковые запросы были взяты с сайта http://wordstat.yandex.ru.

Запросы -

```
[
  "спорт экспресс",
  "виды спорта",
  "активный отдых",
  "хоккеная площадка",
  "трансфер кхл",
  "фигурное катание",
  "профессиональный бокс",
  "боевые искусства",
  "кхл",
  "спортивный клуб",
  "нхл",
  "физическая культура и спорт",
  "лучшие футболисты мира",
  "зимний спорт",
  "зимняя олимпиада",
  "кровавый спорт",
  "газета спорт",
  "министерство спорта",
  "зимние виды спорта",
  "федерация спорта",
  "мастер спорта",
  "олимпийский спорт",
  "команды кхл",
  "конькобежный спорт",
  "чемпионат мира по самбо",
  "лыжный спорт",
  "гиревой спорт",
  "водные виды спорта",
  "самбо",
```

```
  "летняя универсиада"
]
```


Часть запросов достаточно проста и естественна для категории спорт, и предполагаемый вывод сильно совпадает с поисковой выдачей, соответственно имеет высокие оценки. Но есть некоторые несвойственные категории спорт запросы или неоднозначные слова.

Например, достаточно интересный вывод был для запроса «активный отдых». Google справился с этим запросом и выдал страницы, которые содержат информацию о спортивном туризме, видах туризма и прочее. А вот поиск Wiki предложил статьи, которые намного меньше соответствуют ожиданиям и больше описывали туризм, как отдых в различных местах. Так же интересным запросом был «кровавый спорт», который предполагал информацию об одноименной серии фильмов. И здесь же опять Google отработал намного лучше, так как сначала выдача содержала статьи о фильме, а затем об актерах, Wiki же выдала все вперемешку, к тому же были статьи очень отдаленно относящиеся к запросу. К разряду неоднозначных запросов относился запрос «самбо», который обозначает не только вид единоборств, и здесь обе системы выдали достаточно релевантную выдачу.

Подробную информацию об оценках различных метрик можно найти в Git-репозитории(https://github.com/ivadin/search_MAI), а в отчете приведу средние показатели по метрикам обоих поисковых систем.


```
{
      "search_sys": "qoogle",
      "P@1": 1.0,
      "P@3": 1.0,
      "P@5": 1.0,
      "DCG@1": 4.896551724137931,
      "DCG@3": 9.975266464532151,
      "DCG@5": 13.657926517972932,
      "NDCG@1": 0.9793103448275862,
      "NDCG@3": 0.6650177643021434,
      "NDCG@5": 0.5463170607189172,
      "ERR@1": 0.9793103448275862,
      "ERR@3": 0.6650177643021434,
      "ERR@5": 0.5463170607189172
  },
  {
      "search_sys": "wiki",
      "P@1": 0.9655172413793104,
      "P@3": 0.9195402298850575,
      "P@5": 0.9195402298850575,
      "DCG@1": 4.551724137931035,
      "DCG@3": 9.223221721921302,
      "DCG@5": 12.284175920616793,
      "NDCG@1": 0.9103448275862069,
      "NDCG@3": 0.6148814481280869,
      "NDCG@5": 0.4913670368246717,
```

```
        "ERR@1": 0.9103448275862069,
        "ERR@3": 0.6148814481280869,
        "ERR@5": 0.4913670368246717
    }
```


Вывод

В ходе выполнения ЛР были изучены и сравнены поисковые выдачи Wikipedia и Google с ограничением по сайту. Если абстрагироваться от показателей метрик, которые выше для Google, то в субъективной оценке также можно выделить поиск Google, который выдает намного более релевантную выдачу, однако поиск самой Wikipedia также работает достаточно качественно и в большинстве случаев не сильно уступает Google.