**Московский авиационный институт  
(национальный исследовательский университет)**

Факультет прикладной математики и физики  
Кафедра вычислительной математики и программирования

**Лабораторная работа №2  
по курсу «Информационный поиск»**

Оценка качества поиска

|  |  |
| --- | --- |
|  | **Студент:** Рожлейс И.А. **Группа:** М8О-108М-18  **Преподаватель:** Калинин А.Л.  **Дата:** 01.05.2019 |

Москва, 2019

**Цель работы**

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.

Для оценки качества необходимо придумать 30 запросов, отражающих интересы пользователей или, если есть доступ к настоящим запросам пользователей, то выбрать репрезентативную подборку.

Проведите анализ результатов оценки качества. Какие у какой поисковой системы сильные и слабые стороны? Как можно бороться с недостатками, что можно сделать, чтобы улучшить качество?

**Ход работы**

Для начала проведём оценку качества найденного корпуса документов. Для этого посчитаем P-показатель (precision) корпуса документов который определяется как доля релевантных документов к общему количеству документов:

Корпус состоит из 65 тыс. документов. Провести полную оценку качества сложно поэтому ограничимся выборкой из 50 случайных документов. P-показатель корпуса документов оказался равным 0.74, иными словами статистически, лишь 7.4 документа из 10 относятся к выбранной теме «Биология». Такой результат можно объяснить тем что некоторые документы представляют собой смежные и пересекающиеся темы, такие как: химия, биография учёных, описание мифических существ из различных произведений художественной литературы, заповедники и географические местности. Суждение о релевантности документа также является субъективным и поэтому может различаться не только между экспертами, но также могут быть неоднозначно расценены со стороны одного эксперта. Из всего выше сказанного можно сделать заключение что, либо тема корпуса была выбрана слишком общая и её можно разделить на более подробные темы, например, «Учёные биологи», «Животные», «Органические соединения» и т.д., либо было проведено слишком глубокое извлечение документов.

Далее оценим качество поиска Википедии и Google по метрикам качества ранжирования CG (Cumulative Gain), DCG (Discontinued Cumulative Gain), NDCG (Normalized Discontinued cumulative Gain) и ERR (Expected Reciprocal Rank).

Метрика DCG является расширением метрики CG, которая представляет собой сумму рангов первых N документов поискового запроса:

где r – ранг документа.

Метрика CG не учитывает перестановку документов в поисковом запросе и имеет одинаковое значение даже если более релевантный документ находится в конце списка поискового запроса. Для того чтобы учесть положение документа в списке поискового запроса применяется метрика DCG:

Так релевантность документа в начале списка даёт больший вклад в значение метрики, чем релевантность документа в конце списка.

Метрика NDCG нормализует метрику DCG таким образом, чтобы значения DCG можно было сравнивать между собой:

где IDCG некоторое идеальное значение той же метрики DCG. Идеальное значение высчитывается, как если бы список был отсортирован в убывающем порядке, когда релевантные документы находятся в начале списка, а нерелевантный в конце.

Метрика ERR учитывает вероятность того, насколько найденный документ релевантный и соответствует потребностям пользователя:

где N количество документов в списке, а R есть преобразование оценки релевантности документа в вероятность релевантности и задаётся формулой:

Шкала релевантности была выбрана от 0 до 3. Был произведён поиск документов по 30 различным запросам и была составлена таблица релевантности найденных документов.

|  |  |  |
| --- | --- | --- |
| **Запрос** | **Википедия** | **Яндекс** |
| Куда пропали тараканы | 3, 0, 0, 0, 0 | 3, 2, 2, 0, 0 |
| Что находится в пирамидах | 3, 0, 3, 3, 2 | 3, 3, 2, 3, 2 |
| Возможна ли гроза без дождя | 3, 1, 0, 0, 0 | 3, 3, 2, 1, 1 |
| Кто основал Гугл | 3, 2, 0, 0, 0 | 3, 3, 2, 3, 0 |
| Кто основал Википедию | 3, 1, 1, 0, 0 | 3, 3, 3, 1, 0 |
| Что такое шаровая молния | 3, 0, 3, 2, 0 | 3, 3, 3, 1, 0 |
| UNIX-подобные операционные системы | 3, 2, 1, 2, 1 | 3, 3, 3, 1, 2 |
| Краткий пересказ библии | 0, 0, 0, 0, 1 | 3, 3, 2, 1, 2 |
| Фильмы с Брюсом Уиллисом | 3, 3, 3, 3, 3 | 3, 3, 3, 3, 3 |
| Самые странные запросы | 0, 0, 0, 0, 0 | 1, 0, 0, 0, 0 |
| Кто такие Бурбоны | 0, 3, 3, 3, 3 | 3, 3, 0, 3, 2 |
| Почему земля плоская | 3, 1, 1, 0, 0 | 3, 3, 3, 3, 0 |
| Что значит быть в тренде | 1, 0, 1, 0, 1 | 3, 3, 0, 0, 3 |
| Что такое ирония | 3, 1, 0, 2, 1 | 3, 3, 2, 3, 2 |
| Как быстро бегать | 2, 0, 0, 0, 0 | 0, 0, 2, 0, 0 |
| Самые странные праздники | 0, 0, 0, 1, 0 | 3, 0, 0, 0, 1 |
| Где живут инопланетяне | 1, 3, 1, 1, 0 | 3, 3, 3, 3, 3 |
| Происхождение названия Белая Русь | 2, 3, 0, 0, 0 | 3, 1, 1, 3, 0 |
| Лучшие произведения Роберта Шекли | 3, 3, 1, 3, 2 | 3, 3, 3, 3, 3 |
| Откуда прилетают кометы | 0, 0, 1, 0, 0 | 3, 3, 3, 2, 3 |
| Самые удивительные места | 0, 0, 0, 0, 0 | 3, 0, 0, 2, 1 |
| Как появилась нефть | 1, 0, 0, 0, 0 | 3, 3, 3, 1, 3 |
| Как подключить интернет | 3, 1, 1, 1, 1 | 3, 3, 3, 2, 3 |
| Зачем бобру хвост | 3, 3, 2, 2, 0 | 3, 3, 2, 0, 0 |
| Когда возникло христианство | 3, 3, 1, 2, 1 | 3, 3, 3, 3, 3 |
| Могут ли расти ананасы в России | 0, 0, 0, 0, 0 | 2, 0, 2, 2, 0 |
| Существуют ли кварки | 3, 3, 3, 3, 3 | 3, 3, 3, 3, 3 |
| Кто придумал математику | 0, 1, 2, 2, 3 | 3, 3, 2, 3, 2 |
| Чем закончились мстители | 1, 1, 2, 3, 2 | 3, 2, 1, 2, 1 |
| Можно ли отравиться картошкой | 0, 0, 0, 0, 0 | 3, 3, 1, 0, 0 |

**Табл.1.** Релевантность документов поискового запроса.

Для автоматизации вычислений функции метрик были реализованы в MATLAB.

****

Далее была произведена оценка вышеописанных метрик качества на различных уровнях.

Из вышепредставленных таблиц видно, что поисковые машины плохо работают на абстрактных понятиях, таких как например «клетка» (таблицы 3 и 4), которые имеют определение во многих областях, что логично, так как поисковая машина не может определить намерение пользователя искать информацию в заранее известной лишь пользователю области знаний. Имеет место также обратное утверждение: поисковые машины работают очень хорошо если искомый термин имеет определение лишь в одной или нескольких областях (таблицы 5 и 6).

**Заключение**

В работе были оценены различные метрики качества информационного поиска. Была оценена метрика качества корпуса документов.

**Источники**

[1] Olivier Chapelle, Donald Metzler, Ya Zhang, and Pierre Grinspan. Expected Reciprocal Rank for Graded Relevance [Электронный ресурс] // URL: <http://olivier.chapelle.cc/pub/err.pdf>

[2] Chapelle, Metzler, Zhang, Grinspan (2009) Expected Reciprocal Rank for Graded Relevance [Электронный ресурс] // URL: <https://lingpipe-blog.com/2010/03/09/chapelle-metzler-zhang-grinspan-2009-expected-reciprocal-rank-for-graded-relevance>