**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:** 17.03.2019 |

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, 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.

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 – это слишком много чтобы проводить полную оценку 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 (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 собственного поиска Википедии, а также сторонних поисковиков на примере Google и Яндекса. В качестве метрик качества оценим метрики качества ранжирования DCG (Discontinued Cumulative Gain) и NDCG (Normalized Discontinued cumulative Gain).

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.

Наконец 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 ранжирован в строгом соответствии с релевантностью документов.

Шкала релевантности была выбрана как содержащая три значения: нерелевантный документ (0), документ смежной дисциплины (1) и релевантный документ (2).

Далее был произведён поиск документов по нескольким ключевым словам и вычислены метрик DCG и NDCG для каждого запроса на разных уровнях.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **4** | **5** |
| **Википедия** | 2 | 2 | 2 | 2 | 1 |
| **Google** | 2 | 1 | 2 | 2 | 1 |
| **Яндекс** | 2 | 2 | 1 | 1 | 2 |

**Табл.1.** Таблица релевантности запроса «фотосинтез».

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **N=1** | | **N=3** | | **N=5** | |
| **DCG** | **NDCG** | **DCG** | **NDCG** | **DCG** | **NDCG** |
| **Википедия** | 2.00 | 1.00 | 4.26 | 1.00 | 5.51 | 1.00 |
| **Google** | 2.00 | 1.00 | 3.63 | 0.97 | 4.88 | 0.96 |
| **Яндекс** | 2.00 | 1.00 | 3.76 | 1.00 | 4.97 | 0.98 |

**Табл.2.** Метрики запроса «фотосинтез».

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **4** | **5** |
| **Википедия** | 2 | 0 | 2 | 2 | 0 |
| **Google** | 2 | 0 | 0 | 2 | 0 |
| **Яндекс** | 2 | 0 | 0 | 2 | 2 |

**Табл.3.** Таблица релевантности запроса «клетка».

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **N=1** | | **N=3** | | **N=5** | |
| **DCG** | **NDCG** | **DCG** | **DCG** | **DCG** | **NDCG** |
| **Википедия** | 2.00 | 1.00 | 3.00 | 0.92 | 3.86 | 0.91 |
| **Google** | 2.00 | 1.00 | 2.00 | 1.00 | 2.86 | 0.88 |
| **Яндекс** | 2.00 | 1.00 | 2.00 | 1.00 | 3.64 | 0.85 |

**Табл.4.** Метрики запроса «клетка».

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **4** | **5** |
| **Википедия** | 2 | 2 | 2 | 1 | 2 |
| **Google** | 2 | 2 | 2 | 0 | 2 |
| **Яндекс** | 2 | 2 | 2 | 2 | 2 |

**Табл.5.** Таблица релевантности запроса «организм».

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **N=1** | | **N=3** | | **N=5** | |
| **DCG** | **NDCG** | **DCG** | **DCG** | **DCG** | **DCG** |
| **Википедия** | 2.00 | 1.00 | 4.26 | 1.0 | 5.47 | 0.99 |
| **Google** | 2.00 | 1.00 | 4.26 | 1.0 | 5.04 | 0.98 |
| **Яндекс** | 2.00 | 1.00 | 4.26 | 1.0 | 5.90 | 1.00 |

**Табл.6.** Метрики запроса «организм».

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 с учётом предметной области «Биология». Также была оценена метрика качества корпуса документов.