Московский Авиационный Институт
(Национальный исследовательский университет)

Институт №8 «Информационные технологии и прикладная математика»
Кафедра вычислительной математики и программирования

Лабораторная работа №2
по курсу «Информационный поиск»

Студент: Зайцев Н.В.
группа М8О-208М-20

Преподаватель: Кухтичев А.А.

Москва, 2021

Лабораторная работа № 2

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.

Ход работы

В качестве исследуемых поисковых систем были выбраны Google и Mail с ограничением по сайту Википедия, а также встроенный поиск Википедия. Далее проводится расчет метрик.

Метрики поиска Google:

Запрос	P@1	P@3	P@5	DCG@1	DCG@3	DCG@5
поль сизан	1	1	1	4	5,5	6,576498
звездная ночь	1	1	1	4	5	6,189645
кто написал завтрак на траве	1	1	0,8	4	6	5,029086
репин не ждали	1	1	0,8	4	4,5	4,642234
богатыри	1	1	1	1	3	3,868528
фрески рафаэля	1	1	1	4	5	6,576498
сколько грачей на картине грачи прилетели	1	0,333333	0,2	4	2	1,547411
кто расписал сикстинскую капеллу	1	1	1	4	5,5	5,802792
лунная ночь над днепром автор	1	1	1	4	4	5,802792
эпоха возрождения	1	1	1	4	4,5	5,029086
потолок исакиевского собора	1	0,666667	0,4	4	2,5	1,934264
страшный суд картина	1	1	1	4	5	5,802792
в каком жанре писал дали	1	0,666667	0,4	4	2,5	1,934264
импрессионизм это	1	0,666667	0,6	4	3,5	3,868528
вторая мировая война в мировой живописи	1	0,666667	0,8	3	3,5	5,029086
монализа особенности полотна	1	0,666667	0,6	4	3,5	3,481675
картина репина приплыли кто автор	1	0,666667	0,6	4	2,5	2,321117
беллерофонт в походе против химеры русский музей сколько эскизов	1	0,666667	0,4	4	2,5	1,934264
волна айвазовского	1	1	1	4	4,5	4,255381

пинакотека ватикана	1	1	1	4	5,5	6,576498
мадона с младенцем	1	1	1	4	6	7,350203
сколько картин написал даВинчи	1	1	1	4	5	5,415939
итальянская живопись ренессанса	1	1	1	4	6	7,737056
картины с венерой	1	1	1	4	5	6,963351
что такое русский авангард	1	1	1	4	5,5	6,963351
руско турецкая война в живописи	1	1	1	3	5	6,189645
коллекция полотен эрмитажа	1	1	1	3	4,5	5,802792
экспозиция лувра	1	1	1	4	5	5,802792
василий поленов биография	1	0,666667	0,8	4	4	3,868528
кто автор витязя на распутье	1	0,333333	0,4	4	2	2,70797

Запрос	**NDCG@1**	**NDCG@3**	**NDCG@5**
поль сизан	1	0,916667	0,85
звездная ночь	1	0,833333	0,842105
кто написал завтрак на траве	1	1	0,722222
репин не ждали	1	0,818182	0,666667
богатыри	0,25	0,545455	0,588235
фрески рафаэля	1	0,909091	1
сколько грачей на картине грачи прилетели	1	0,363636	0,235294
кто расписал сикстинскую капеллу	1	1	0,9375
лунная ночь над днепром автор	1	0,8	0,9375
эпоха возрождения	1	0,9	0,866667
потолок исакиевского собора	1	0,5	0,333333
страшный суд картина	1	1	1
в каком жанре писал дали	1	0,5	0,333333
импрессионизм это	1	0,7	0,666667
вторая мировая война в мировой живописи	0,75	0,777778	0,928571
монализа особенности полотна	1	0,777778	0,692308
картина репина приплыли кто автор	1	0,555556	0,461538
беллерофонт в походе против химеры русский музей сколько эскизов	1	0,555556	0,384615
волна айвазовского	1	1	0,916667
пинакотека ватикана	1	1	1

мадона с младенцем	1	0,909091	0,9375
сколько картин написал даВинчи	1	1	0,85
итальянская живопись ренессанса	1	1	0,722222
картины с венерой	1	0,9	0,85
что такое русский авангард	1	0,818182	0,692308
руско турецкая война в живописи	0,75	1	0,866667
коллекция полотен эрмитажа	1	1	1
экспозиция лувра	1	0,777778	0,85
василий поленов биография	1	0,833333	0,666667
кто автор витязя на распутье	1	1	0,842105

Метрики поиска Mail:

Запрос	**P@1**	**P@3**	**P@5**	**DCG@1**	**DCG@3**	**DCG@5**
поль сизан	1	0,666667	0,4	4	2,5	1,934264
звездная ночь	1	0,666667	0,4	4	3,5	2,70797
кто написал завтрак на траве	1	0,666667	0,6	4	4	3,481675
репин не ждали	1	0,666667	0,4	4	4	3,094822
богатыри	1	1	0,6	4	4	3,094822
фрески рафаэля	1	1	0,8	4	5	5,029086
сколько грачей на картине грачи прилетели	1	0,666667	0,4	4	3	2,321117
кто расписал сикстинскую капеллу	1	1	1	4	5	5,802792
лунная ночь над днепром автор	1	0,333333	0,6	4	2	3,868528
эпоха возрождения	0	0,333333	0,2	0	2	1,547411
потолок исакиевского собора	1	0,666667	0,4	3	3,5	2,70797
страшный суд картина	1	0,666667	0,8	4	4	5,802792
в каком жанре писал дали	0	0	0	0	0	0
импрессионизм это	0	0,333333	0,2	0	2	1,547411
вторая мировая война в мировой живописи	0	0	0	0	0	0
монализа особенности полотна	1	0,666667	0,4	4	3,5	2,70797
картина репина приплыли кто автор	1	1	0,6	2	4,5	3,481675
беллерофонт в походе против химеры русский музей сколько эскизов	1	0,666667	0,4	2	3	2,321117
волна айвазовского	1	0,666667	0,8	1	2,5	5,029086

пинакотека ватикана	1	0,666667	0,8	4	4	3,868528
мадона с младенцем	1	1	1	4	6	7,737056
сколько картин написал даВинчи	1	1	0,8	2	5	4,642234
итальянская живопись ренессанса	1	0,666667	0,6	4	4	4,642234
картины с венерой	1	0,666667	0,6	4	4	4,642234
что такое русский авангард	1	0,666667	0,4	4	3,5	2,70797
руско турецкая война в живописи	1	0,333333	0,6	4	2	3,868528
коллекция полотен эрмитажа	1	0,666667	0,6	3	3,5	4,255381
экспозиция лувра	1	1	1	4	4,5	4,255381
василий поленов биография	1	0,666667	0,6	4	3	3,094822
кто автор витязя на распутье	1	0,333333	0,2	4	2	1,547411

Запрос	**NDCG@1**	**NDCG@3**	**NDCG@5**
поль сизан	1	0,416667	0,25
звездная ночь	1	0,7	0,466667
кто написал завтрак на траве	1	0,8	0,6
репин не ждали	1	0,8	0,615385
богатыри	1	0,888889	0,615385
фрески рафаэля	1	0,888889	0,783333
сколько грачей на картине грачи прилетели	1	0,75	0,5
кто расписал сикстинскую капеллу	1	1	1
лунная ночь над днепром автор	1	0,5	0,909091
эпоха возрождения	0	0,5	0,363636
потолок исакиевского собора	0,75	0,875	0,7
страшный суд картина	1	1	0,75
в каком жанре писал дали	0	0	0
импрессионизм это	0	0,571429	0,444444
вторая мировая война в мировой живописи	0	0	0
монализа особенности полотна	1	1	0,875
картина репина приплыли кто автор	0,5	0,857143	0,909091
беллерофонт в походе против химеры русский музей сколько эскизов	0,5	0,857143	0,75
волна айвазовского	0,25	0,833333	0,980241
пинакотека ватикана	1	0,571429	0,615385

мадона с младенцем	0,666667	0,8	0,980241
сколько картин написал даВинчи	0,666667	1	0,909091
итальянская живопись ренессанса	1	0,6	1
картины с венерой	1	1	1
что такое русский авангард	1	0,75	0,875
руско турецкая война в живописи	1	1	0,909091
коллекция полотен эрмитажа	0	0,75	0,783333
экспозиция лувра	0	0,875	0,783333
василий поленов биография	0	0	0
кто автор витязя на распутье	0	0	0

Метрики поиска Википедии:

Запрос	**P@1**	**P@3**	**P@5**	**DCG@1**	**DCG@3**	**DCG@5**
поль сизан	0	0	0	0	0	0
звездная ночь	1	0,666667	0,4	4	3,5	2,70797
кто написал завтрак на траве	1	0,666667	0,6	4	4	3,481675
репин не ждали	1	0,666667	0,6	4	4	3,868528
богатыри	1	1	0,6	3	5	3,868528
фрески рафаэля	1	1	1	4	6	6,576498
сколько грачей на картине грачи прилетели	1	0,333333	0,2	4	2	1,547411
кто расписал сикстинскую капеллу	1	0,666667	0,8	4	4	5,029086
лунная ночь над днепром автор	1	1	0,6	4	4,5	3,481675
эпоха возрождения	1	0,666667	0,4	4	4	3,094822
потолок исакиевского собора	0	0,333333	0,2	0	0,5	0,386853
страшный суд картина	1	1	1	3	5,5	6,576498
в каком жанре писал дали	0	0	0	0	0	0
импрессионизм это	0	0,333333	0,4	0	1	2,321117
вторая мировая война в мировой живописи	0	0	0,2	0	0	1,160558
монализа особенности полотна	0	0	0	0	0	0
картина репина приплыли кто автор	0	0	0	0	0	0
беллерофонт в походе против химеры русский музей сколько эскизов	1	0,333333	0,2	4	2	1,547411
волна айвазовского	1	1	1	4	4,5	4,255381

пинакотека ватикана	1	0,666667	0,6	4	3,5	3,868528
мадона с младенцем	1	1	1	2	5	6,963351
сколько картин написал даВинчи	0	0	0	0	0	0
итальянская живопись ренессанса	1	1	1	4	5	6,963351
картины с венерой	1	1	1	4	6	7,737056
что такое русский авангард	1	1	0,8	4	5,5	5,415939
руско турецкая война в живописи	0	0	0	0	0	0
коллекция полотен эрмитажа	1	1	0,6	4	5	3,868528
экспозиция лувра	1	0,333333	0,4	4	2	2,321117
василий поленов биография	1	0,666667	0,4	4	3,5	2,70797
кто автор витязя на распутье	0	0	0	0	0	0

Запрос	**NDCG@1**	**NDCG@3**	**NDCG@5**
поль сизан	0	0	0
звездная ночь	1	0,583333	0,388889
кто написал завтрак на траве	1	0,727273	0,5
репин не ждали	1	0,727273	0,588235
богатыри	0,75	1	0,588235
фрески рафаэля	1	0,25	0,9
сколько грачей на картине грачи прилетели	1	0,4	0,307692
кто расписал сикстинскую капеллу	1	0,8	0,888889
лунная ночь над днепром автор	1	1	0,9
эпоха возрождения	1	0,888889	0,8
потолок исакиевского собора	0	0,125	0,1
страшный суд картина	0,75	1	1
в каком жанре писал дали	0	0	0
импрессионизм это	0	0,25	0,666667
вторая мировая война в мировой живописи	0	0	0,375
монализа особенности полотна	0	0	0
картина репина приплыли кто автор	0	0	0
беллерофонт в походе против химеры русский музей сколько эскизов	0,666667	1	0,666667
волна айвазовского	0,666667	0,8	1
пинакотека ватикана	1	1	0,727273

мадона с младенцем	1	0	1
сколько картин написал даВинчи	0	0	0
итальянская живопись ренессанса	1	1	0,8
картины с венерой	0	0	0,984132
что такое русский авангард	1	1	0,888889
руско турецкая война в живописи	0	0	0
коллекция полотен эрмитажа	1	0,583333	1
экспозиция лувра	1	0,583333	0,388889
василий поленов биография	1	1	0
кто автор витязя на распутье	0	0	0

Подводя итог, можно сделать следующие выводы по метрикам для поисковых систем:

	Google	**Mail**	**Википедия**
P@1	1	0,866667	0,666667
P@3	0,866667	0,644444	0,544444
P@5	0,826667	0,54	0,466667
DCG@1	3,8	3,1	2,533333
DCG@3	4,283333	3,316667	2,866667
DCG@5	4,900136	3,39141	2,991662
NDCG@1	0,958333	0,644444	0,594444
NDCG@3	0,823047	0,686164	0,490614
NDCG@5	0,75469	0,645592	0,515315

Вывод

Проанализировав метрики качества поиска для систем Google, Mail и внутреннего поиска Википедии, можно сделать вывод о том, что внутренний поиск Википедии во многом уступает другим изученным системам. Особенно это видно на запросах, требующих понимания того, что именно пользователь хочет видеть в выдаче. Поиск Википедии в этом случае выдает статьи, основываясь на вхождении слов запроса, из-за чего результат часто не является релевантным.

Однако, в целом, когда запрос нацелен на поиск чего-то конкретного, например, «Прогноз погоды» или «Торрент», метрики показывают незначительное отличие между поисковыми системами. Но при сложных запросах поиск Google выигрывает.