**PROJECT REPORT**

Comparison of execution time among different methods for different queries.

Query 1: "KASHMIR BRITISH"

Query 2: "A GOVERNMENT POLICEMAN HAPPENED IN ON THE PROCEEDINGS AND REPORTED"

Query 3: "PEKING KREMLIN MOSCOW"

Query 4: "PULSATING WEEK CHINA LAST"

Query 5: "COMMUNISM SCHISM GRAVEST MOSCOW"

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | EQ | CL | IE | CP |
| Query 1 | 0.0040025 | 0.01382398 | 0.0020031 | 0.0020010 |
| Query 2 | 0.0160136 | 0.01101160 | 0.0 | 0.0020017 |
| Query 3 | 0.0148167 | 0.01501011 | 0.0150103 | 0.0055704 |
| Query 4 | 0.0155415 | 0.01554346 | 0.0 | 0.0155439 |
| Query 5 | 0.0155434 | 0.02205109 | 0.0 | 0.0010992 |

Time isn’t a major factor as the number of documents in our experiment is limited. It becomes a factor when we are dealing with big data. With exact query as the benchmark, it is found that Index Elimination method retrieves the right results in the shortest amount of time. The ranking order of the results may be slightly different compared to Index elimination.

Execution time isn’t a factor in our case as the system configuration is very capable of dealing with the execution of the code for the experiments.

In case of champion lists, I have chosen the value for r documents to be square root of collection length

The value of k is provided before execution time and the r documents are sorted in decreasing order of their weights/TF-IDF values.

A better measure of the ranked retrieval results is the Precision, Mean Average Precision to be precise.

The following are the experiment results from three different inexact methods run for 5 different queries.

I have calculated precision and recall for three inexact methods based on the presence of exact query results in their results. If the top 10 results of these methods are in the top 10 results of the exact method, then they are relevant, otherwise they are non-relevant.

Precision = Total Relevant Documents / Retrieved Documents

Recall = Relevant Documents retrieved / Total Relevant Documents

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | EQ | CL | IE | CP |
|  |  |  |  |  |
| Q1 | Text-184.txt | Text-184.txt | Text-184.txt | Text-184.txt |
|  | Text-63.txt | Text-63.txt | Text-63.txt | Text-7.txt |
|  | Text-7.txt | Text-7.txt | Text-7.txt | Text-346.txt |
|  | Text-346.txt | Text-346.txt | Text-346.txt | Text-46.txt |
|  | Text-36.txt | Text-10.txt |  |  |
|  | Text-10.txt | Text-254.txt |  |  |
|  | Text-254.txt | Text-61.txt |  |  |
|  | Text-61.txt | Text-247.txt |  |  |
|  | Text-109.txt | Text-339.txt |  |  |
|  | Text-247.txt | Text-117.txt |  |  |
|  |  |  |  |  |
|  | AVERAGE PRECISION----🡪 | 0.8 | 1 | 0.75 |
|  | RECALL---🡪 | 0.8 | 0.4 | 0.3 |
| Q2 | Text-143.txt | Text-143.txt | Text-143.txt | Text-16.txt |
|  | Text-213.txt | Text-213.txt |  | Text-197.txt |
|  | Text-256.txt | Text-256.txt |  | Text-235.txt |
|  | Text-141.txt | Text-183.txt |  |  |
|  | Text-183.txt | Text-141.txt |  |  |
|  | Text-69.txt | Text-69.txt |  |  |
|  | Text-16.txt | Text-16.txt |  |  |
|  | Text-333.txt | Text-333.txt |  |  |
|  | Text-224.txt | Text-83.txt |  |  |
|  | Text-83.txt | Text-224.txt |  |  |
|  |  |  |  |  |
|  | AVERAGE PRECISION----🡪 | 1 | 1 | 0.34 |
|  | RECALL---🡪 | 1 | 0.1 | 0.1 |
| Q3 | Text-111.txt | Text-296.txt | Text-363.txt | Text-385.txt |
|  | Text-216.txt | Text-385.txt | Text-111.txt | Text-353.txt |
|  | Text-385.txt | Text-216.txt | Text-385.txt | Text-104.txt |
|  | Text-353.txt | Text-275.txt | Text-353.txt | Text-401.txt |
|  | Text-296.txt | Text-104.txt | Text-216.txt | Text-307.txt |
|  | Text-275.txt | Text-289.txt | Text-296.txt | Text-279.txt |
|  | Text-104.txt | Text-145.txt | Text-18.txt | Text-363.txt |
|  | Text-401.txt | Text-307.txt | Text-374.txt | Text-417.txt |
|  | Text-289.txt | Text-279.txt | Text-104.txt | Text-18.txt |
|  | Text-145.txt | Text-353.txt | Text-275.txt | Text-8.txt |
|  |  |  |  |  |
|  | AVERAGE PRECISION----🡪 | 0.8 | 0.7 | 0.4 |
|  | RECALL---🡪 | 0.8 | 0.7 | 0.4 |
| Q4 | Text-222.txt | Text-180.txt | Text-180.txt | Text-363.txt |
|  | Text-180.txt | Text-222.txt | Text-222.txt | Text-8.txt |
|  | Text-363.txt | Text-363.txt |  | Text-31.txt |
|  | Text-8.txt | Text-8.txt |  | Text-385.txt |
|  | Text-31.txt | Text-31.txt |  | Text-199.txt |
|  | Text-346.txt | Text-346.txt |  | Text-18.txt |
|  | Text-385.txt | Text-385.txt |  | Text-401.txt |
|  | Text-420.txt | Text-199.txt |  | Text-104.txt |
|  | Text-199.txt | Text-374.txt |  | Text-417.txt |
|  | Text-374.txt | Text-18.txt |  | Text-279.txt |
|  |  |  |  |  |
|  | AVERAGE PRECISION----🡪 | 0.9 | 1 | 0.5 |
|  | RECALL---🡪 | 0.9 | 0.2 | 0.5 |
| Q5 | Text-289.txt | Text-289.txt | Text-289.txt | Text-279.txt |
|  | Text-279.txt | Text-279.txt | Text-279.txt | Text-190.txt |
|  | Text-332.txt | Text-332.txt |  | Text-332.txt |
|  | Text-341.txt | Text-290.txt |  | Text-187.txt |
|  | Text-290.txt | Text-295.txt |  | Text-265.txt |
|  | Text-295.txt | Text-341.txt |  |  |
|  | Text-146.txt | Text-269.txt |  |  |
|  | Text-269.txt | Text-265.txt |  |  |
|  | Text-265.txt | Text-81.txt |  |  |
|  | Text-81.txt | Text-60.txt |  |  |
|  |  |  |  |  |
|  | AVERAGE PRECISION----🡪 | 0.9 | 1 | 0.6 |
|  | RECALL---🡪 | 0.9 | 0.2 | 0.3 |
|  |  |  |  |  |

Please refer to the Experiments.xlsx for clearer representation of the experiments conducted.

From the above data,

Mean Average Precision for Cluster Pruning: (0.75+0.33333+0.4+0.5+0.6)/5 = 0.51667

Mean Average Precision for Champion Lists: (1+1+0.975+1+1)/5 = 0.995

Mean Average Precision for Index Elimination: (1+1+0.702+1+1)/5 = 0.941

Average Precision for Cluster Pruning: (0.6+0.3+0.4+0.4+0.7)/5 = 0.55

Average Precision for Champion Lists: (0.8+0.8+1+0.9+0.9)/5 = 0.88

Average Precision for Index Elimination: (1+1+0.7+1+1)/5 = 0.94

Average Recall for Cluster Pruning: (0.6+0.3+0.4+0.4+0.7)/5 = 0.32

Average Recall for Champion Lists: (0.8+0.8+1+0.9+0.9)/5 = 0.88

Average Recall for Index Elimination: (1+1+0.7+1+1)/5 = 0.32

Tradeoff between Precision and Recall, the F-measure:

F-measure = 2PR/P+R

F for Cluster Pruning: = 0.41

F for Champion Lists: = 0.88

F for Index Elimination: = 0.48

I can safely infer from my experiments that Champion Lists has a proven record of getting the right/relevant documents quicker than most other inexact retrieval methods. Index elimination comes a close second.

Factors that have impact on the performance of this program are the number of documents. If the number of documents is doubled, creating clusters and building index. The combined execution time of these two components.