搜索引擎 PA_2

——实验报告

闫力敏*

1 概述

本次实验主要实现一个简单的搜索引擎，然后通过给定的 Query 集进行测试，最后将所得结果采用给定的 DCG，N-Err 以及 Q-Measure 指标评测

2 实验过程

利用 lucene 提供的框架，实现一个简单的搜索引擎，其中通过继承 SimilarityBase 类重载 score 实现自定义排序 BM25

查询区域选取为 title 和 content 两个区域，分别计算出相关的分数之后，按照 3：1 加权计算得到最终的结果

$$BM25_{final} = BM25_{title} * 0.75 + BM25_{content} * 0.25$$

每次查询返回 top20 的规定结果(包括分数),提取出 y_pred,随后利用给定的 ntcir14_test_label.txt 数据文件提取出对应的 y_true，根据要求的评测指标进行计算。

3 实验结果

k	n_dcg	q_measure	n_err
20	0.441	0.329	0.689
10	0.514	0.412	0.686
5	0.692	0.592	0.684

top_N 的结果整体来说随着 N 的减少，DCG，Q-Measure，正相关，且比较明显，而 N-Err 负相关，但差异变化不大，总体来说还是比较符合预期结果的

4 总结

本次实验，总体来说还是比较简单的，主要的收获还是学到了一些比较有用的工具

*清华大学计算机系. 学号：2015011391. 邮编：100084