Query导向的排序方法评价实验

Data Collection

- 根据数据库内容编写query
  - query的数量
  - query的长度
  - query的质量
  - 包含该query的字段
- 根据query得到hits
  - 对hits根据query进行打分（3分制或4分制或5分制）
  - 3-5人对同一数据集进行打分，不同取多数，无多数取平均
  - 给用户呈现哪些内容
  - 质量和相关性评判标准（打分标准）
  - 是否区分质量分数与相关性分数

- 具体流程：
  i. 统计**class&property**字段、**title&notes**字段中出现次数最多的前100个有效term（去除**含数字的或长度为1的**term）；
  ii. 在google dataset search上搜索这些term，取**搜索下拉框推荐的相关查询**作为origin queries（每个term10个，个别term不足10个）；
  iii. 取origin queries中**去掉标点符号**后使用Lucene(8.7.0)中English Analyzer **parse后长度小于等于8**的且**只包含英语和数字**的作为test queries；
  iv. 取test queries中使用Lucene默认评分函数**平均每词每字段得分大于阈值k的hits数多于20的**作为实验用查询；
  v. 取这些查询不同baseline的hits前20集合；
  vi. 将得到的hits随机提供给用户（3人以上）进行打分，根据该数据集与该query的相关程度进行打分。

- property：所有predicate
- class：所有'%rdf-syntax-ns#type%'的predicate指向的object

Pooling

- Setup
  - BM25, TFIDF的计算字段为title, notes, class, property(简单求和 or 加权平均)；
  - FSDM微调class和property字段的权重；
  - BM25, TFIDF, FSDM使用Lucene默认方法检索得到前500个hits后进行rerank；DPR直接对全部数据集文本进行检索；

- Result

labeling guidance

- a dataset is off topic (0) if the information does not satisfy the information need, and should not be listed in the search results from a search engine;
- a dataset is poor (1) if a search engine were to include this in the search results, but it should not be listed at the top;
- a dataset is good (2) if you would expect this dataset to be included in the search results from a search engine;
- a dataset is excellent (3) if you would expect this dataset ranked near the top of the search results from a search engine.

Baseline

- 单一方法排序
  - TF-IDF
  - BM25
  - FSDM
  - DPR
  - PageRank
  - DING
  - DRank (仅根据数据集的度数排序的naive rank)

- 混合方法排序
  - Quality(PageRank、DING、DRank) + Relevance(TF-IDF、BM25、FSDM)
- 多filed与单field
  - 全部field
  - 仅content
  - 仅title和description
  - 仅content、title和description

Research Questions

- Q1 不同方法的效果比较（在单field上对比和在多field上对比）
- Q2 不同field对排序效果的影响（相同方法下不同field上搜索的比较）
- Q3 自主设计方法的效果

Evaluation Metrics

- nDCG@k (k=5,10,15,20*)
- Precision@k
- Recall@k
- F値 or MAP