ICTNET在2012 web track的试验方法

1.建索引 用新的网页解析器，将网页中的广告和spam去掉，抽取了 trec-id,title, url,正文，锚文本 2.锚文本的使用 对锚文本的检索有效的提升了搜索效果（参见ICTNET2011）；锚文本的数据采用Lemor官方提供的锚文本数据集。采用了map-reduce的方法来统计各url中的锚文本 3. 建立索引 1)使用天矶的索引器，对trec-id,url,title,正文，锚文本，spam value进行索引。2)在10台机器上使用天机2进行索引，10个小时就搞定了。 4.排序模型–BM25 使用boolean模型和bm25模型的结合。 在短的域上(title，锚文本)使用OR形式，对长的域用AND形式。最后对所有的备选文档计算bm25的值，然后加和排序。 5.排序模型–learning to rank 采用来RANKBOOST的模型，训练数据是web trec 2009和2010的结果集，经过5-folds的交叉验证后直接用于排序。 6.结合wikipedia结果 用wiki中的结果作为top1，用其他的结果作为补充。 7.结果分析 一共三轮结果。 a，采用原来的文本抽取器，BM25，wiki数据，作为baseline b, 采用新文本抽取器，BM25，WIKI，在ERR@20上比a好，其他都不如a，说明正文提取对于准确率有帮助，但是降低了召回率。 c，采用排序学习方法，效果比a差很多，可能是由于训练样本太少造成。 8.综上 简单的排序模型即可，关键在于如何筛选掉错误的备选。