# 体育新闻搜索平台设计文档

361°检索平台

现代信息检索大作业

成员：

姬强 201428015059068

徐培兴

倪佳志

陈晓旭

题目要求：

定向采集3-4个体育新闻网站，实现这些网站信息的抽取、索引和检索。网页数目不少于10万条。能按相关度、时间、热度(需要自己定义)等属性进行排序，能实现相似新闻的自动聚类。

基本处理流程：

1. 定向爬取数据
2. 分析爬取数据，提取相关字段，写入数据库中
3. 分词，并将分词结果写入数据库中方便下一步处理
4. 构建倒排索引
5. 检索，排序
6. 前台展示

具体说明：

1. 新闻数据的获取以及预处理
   1. 如何爬取的数据

1.2爬取数据统计：

|  |  |  |
| --- | --- | --- |
| 网站 | Type | 新闻数量 |
| 新浪体育 | 2 | 57378 |
| 腾讯体育 | 3 | 26405 |
| 网易体育 | 4 | 16496 |
| 总计： | 100279 条 | |

( SELECT COUNT(id) from htmls where type = 3)

1.3数据处理 （解析 和 分词、停用词过滤）

解析的时候的问题：主要是网页的结构的变化，以网易为例，从06年到14年，网页中标签属性和id的变化次数大概为8次，几乎每年都会发生若干变化。时间格式，有的时候也会发生变化。这些变化都是不可提前预知的，因此在一定程度上影响了数据的获取。同时在网页中还有部分网页是“特殊网页”，比如视频，图片等，这些也需要在解析网站的时候特别判断。

解析的数据域包括： 新闻标题title、新闻的keywords、摘要描述description、新闻时间、新闻热度、新闻url、新闻主体body、

**停用词表**： 由于是体育新闻的检索，除了常用的停用词外，还有如下停用词：体育、比赛

1.4 存储 mysql 各个表的用途

具体表的用途以及各个字段。

1. 倒排索引的构建

三个倒排索引，优先级由高到底

1. 热度较高的词的倒排索引
2. 新闻的标题title，关键词keywords、摘要description的分词结果构建的倒排索引
3. 新闻主体构建的倒排索引

由高到低查询索引，设定阈值Threshold\_pageNum，当在高级别的索引中查询到的文档数量超过阈值的时候，低级别的索引不再查询，只有当用户请求更多的文档的时候才进行查询。

1. 查询结果的排序实现

有多个词的时候：

排序参考的因素：文档中出现的查询中的词项数量、文档中出现的查询中的词的tfidf信息、新闻时间、评论数（热度）

标记：

文档D 长度为|D|（使用文档中词的个数来代表）

查询 Q长度为|Q|（同样也是分词后的词的数量）

文档和查询共有词：D∩Q 词量为 | D∩Q |

TFIDF 为 共有词的tfidf 之和

只要| D∩Q |的值大于0的文档都会返回，如果返回的数量较大，去除那些共有词出现次数小于K的文档。

1. 前台展示的实现
   1. 自动补全实现
   2. 大纲展示
   3. 重点标红
2. 聚类

经验总结：

该项目有四人合力协作完成，锻炼了大家的团队协作能力，加深了对检索系统的了解，基本实现了体育新闻的检索，能够较好的根据查询返回结果。不足之处也很多，由于时间紧迫，在排序函数的设计上还可以更好。还可以通过其他方式，比如分布式来加快检索过程。