# 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 | 新闻数量 | 起始时间 | 截止时间 |
| 新浪体育 | 50，000条 |  |  |
| 搜狐体育 |  |  |  |
| 网易体育 |  |  |  |
| 总计： |  |  |  |

1.3数据处理 （解析 和 分词）

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
2. 聚类的实现
3. 查询结果的排序实现

排序参考的因素， 评分：时间、评论数、类簇的大小

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

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。