# 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。

基本模块：

1. 数据获取和预处理
2. 分词工具
3. 索引构建
4. 向量空间索引
5. 聚类

使用的开源的工具/jar包和用途：

|  |  |
| --- | --- |
| 开源工具/jar包 |  |
| Jsoup |  |
|  |  |
|  |  |

**数据库表的列表和用途**

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

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 | |

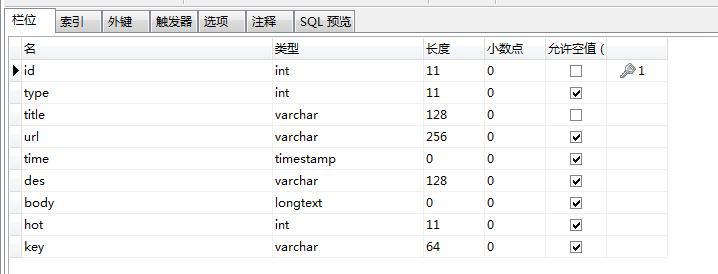
( SELECT COUNT(id) from htmls where type = 3)

1.3数据处理 （解析 和 分词、停用词过滤）

|  |  |
| --- | --- |
| 类名 | 功能 |
| GetInforFormHtml.java | 处理html文件，分析得到数据 |
| HtmlDetailManager.java | 将详细数据写入到数据库中 |
| 问题：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。 | |

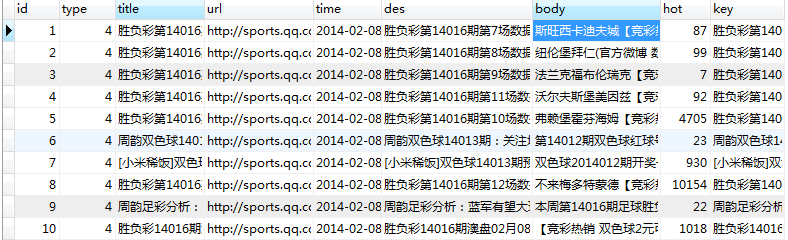
解析的数据域包括： 新闻标题title、网站类型type、新闻的keywords、摘要描述description、新闻时间、新闻热度、新闻url、新闻主体body

|  |  |
| --- | --- |
| 网站类别 | Type |
| *SINA* | 2 |
| *WanyYi* | 3 |
| *Tencent* | 4 |



**网页数据解析后存储表结构**

数据示例：



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

采用组平均凝聚式聚类算法

**输入**： 查询返回的文档列表以及文档的得分

**输出**： 聚类后的按照得分由高到低进行排序的簇列表

处理： 设定了聚类停止阈值：

第一个阈值是簇的数量Math.sqrt(DocNum)

第二个是相似度大小sim\_threshold

上面两个阈值有一个不满足就结束聚类，经过测试，sim\_threshold取值为0.5的时候效果不怎么理想，几乎每个簇都是只有一个到两篇文档，当为0.2的大部分时候返回的聚类的结果也不是很理想，一千篇文档聚出900多个类。 当继续调低阈值的时候，簇的数量不断减少，当小于0.1的时候簇的数量急剧减小，平均情况下1000篇文档大概聚出不到100篇文。为了使得绝大部分聚类后的簇的数量限制在较小的水平上，我们选择0.02作为阈值，500篇文档，簇数量大部分小于50。

降维：给定阈值termThreshold，将小于该值的维度去掉。

当不处理的时候，簇的质心的维度一般维度较高，有的可以达到几千维，当阈值设定为0.01，降低为一般不超过200-300维度，当设定为0.1的时候结果发现聚类后的簇的质心的维度一般最大不超过200维。 由于一篇新闻一般在200-300个词汇这里选择0.01作为阈值。

代码：

|  |  |
| --- | --- |
| 类名 | 用途 |
| BuildDocVectorTable.java | 1. 构建向量空间 2. 将归一化向量写入到数据库中 |
| ClusterDetail.java | 存放一个簇的详细信息   1. 记录簇的标签（这里使用了距离质心最近的文档的标题） 2. 得分 3. 簇中文档列表 4. 簇的质心 |
| Clustering\_GAAC.java | 具体实现聚类的细节：  1. 计算文档相似度  2. 计算簇间的相似度  3. 合并簇  4. 对簇进行排序  5. 计算簇的质心，并选择距离质心最近的文档的标题作为簇的标签 |
| 评价： 总体感觉聚类效果一般，可能是由于检索的结果本身不像有歧义的词那样会产生类别比较明显的结果。 | |

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。