1 数据库课程设计准备

1.1 前言

这里的内容仅作为初期对于系统功能的设想，并未成体系，因此内容稍有凌乱

在正式开发之后会对于这个部分进行正式的整理和规划

1.2 任务书具体描述

1.2.1 题目：基于 AI 的新闻热点聚合及可视化系统

针对互联网几大主要门户网站进行内容爬取，将爬取的内容进行内容解析、分析和统计， 将分析结果数据写入数据库系统，根据热点主题抽取，可视化展示聚合内容。

1.3 功能分析和设计

1. 用户模块 —— **需要一个表保存用户和管理员的信息**
    - 管理员模块
        - 删除、查看用户
        - 添加 or 删除可爬取的目标网站 —— **需要一个表保存可以进行爬取的目标网站**

            > 不能让用户自己写目标网站，避免违规爬取某些网站，同时也避免网站的代码风格不同解析失败爬取到错误信息

        - 对中间任何一步都有插手进行更改的权力

            > 这是联系实际情况，新闻也有压热度，删热词的情况，因此设想出的一个功能，保证平台的新闻某种程度上是能够“不越界”的

    - 用户模块
2. 新闻数据采集模块
    - 新闻数据爬取 —— **需要一个表保存爬取的信息**，原因如下

        > 新闻数据爬取可以考虑每一个小时发一次请求，而不是用户点击一次发一次请求，避免出现用户连续点击造成 ddos 攻击危险
        >
        > 同时如果多用户同时要对网站新闻进行分析，也避免重复的新闻响应，造成资源浪费

3. 新闻分词预处理 —— 分词同样需要表保存

    > 如果用户发现分词中某些内容不是自己所需要的，可以进行删除和修改，以此来做二次的筛选
    >
    > 因此每个用户有自己的一套对所有新闻提取后并可以预处理的分词库（一个人一个，重新提取的时候进行刷新）

4. 新闻数据显示模块

    > 这里给用户两个显示方案

> 1. 查看新闻关键词提取得到的数据，然后直接进行分析
> 2. 拿到原始数据，然后进行一些修改后再进行分析

5. 新闻标签提取，支持管理员二次编辑和处理

> 借由NLP相关算法，由新闻分词库得到对应的一系列新闻标签（这里的标签类似于关键词）

6. 新闻分类，支持管理员二次编辑和处理

> 和上面同理，只不过输入的各个新闻，输出是这个新闻对应的分类

7. 新闻数据支持，支持管理员二次编辑和处理

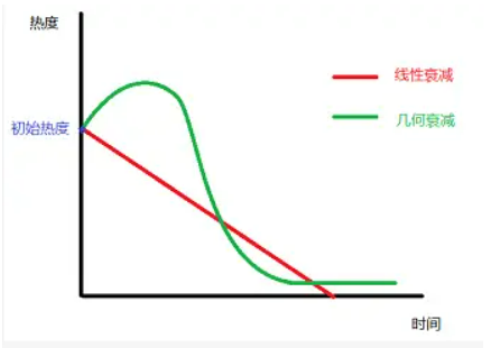
> 编辑功能可以让用户能够人为删除一些错误的干扰信息，或者认为没有用处的新闻

8. 新闻关键热点词统计分析模块

> 下面的这个内容是我对于热度统计算法的一个调查所获得的较好的算法解释，我将会根据这个部分对热度算法进行设计

2.4 热度随时间的衰减不是线性的

由于新闻的强时效性，已经发布的新闻的热度值必须随着时间流逝而衰减，并且趋势应该是衰减越来越快，直至趋近于零热度。换句话说，如果一条新闻要一直处于很靠前的位置，随着时间的推移它必须要有越来越多的用户来维持。



我们要求推荐给用户的新闻必须是24h以内，所以理论上讲，衰减算法必须保证在24h后新闻的热度一定会衰减到很低，如果是线性衰减，当某些新闻突然有大量用户阅读，获得很高的热度分时，可能会持续排名靠前很久，让用户觉得内容更新过慢。

参考牛顿冷却定律，时间衰减因子应该是一个类似于指数函数：

$$T(Time) = e \wedge (k*(T1 - T0))$$

其中T0是新闻发布时间，T1是当前时间。

而由于热度的发展最终是一个无限趋近于零热度的结果，最终的新闻的热度算法也调整为：

$$Score = (S0(Type) + S(Users)) / T(Time)$$

2.1 热度算法基本原理

需要了解的是，热度算法也是需要不断优化去完善的，基本原理：

> 新闻热度分 = 初始热度分 + 用户交互产生的热度分 – 随时间衰减的热度分
>
> Score = S0 + S(Users) – S(Time)

新闻入库后，系统为之赋予一个初始热度值，该新闻就进入了推荐列表进行排序；随着新闻不断被用户点击阅读，收藏，分享等，这些用户行为被视作帮助新闻提升热度，系统需要为每一种新闻赋予热度值；同时，新闻是有较强时效性的内容，因此新闻发布之后，热度必须随着新闻变得陈旧而衰减。

新闻的热度就在这些算法的综合作用下不断变化，推荐列表的排序也就不断变化。

2.2 初始热度不应该一致

上面的算法为每一条入库的新闻赋予了同样的热度值，但在现实使用后发现行不通，例如娱乐类别比文化类别受欢迎程度本身就高很多；或者突发了严重的灾害或事故；或是奥运会期间，体育类别的关注度突然高了起来；而此时如果还是每条新闻给同样的热度就不能贴合实际了。

解决办法就是把初始热度设置为变量：

（1）按照新闻类别给予新闻不同的初始热度，让用户关注度高的类别获得更高的初始热度分，从而获得更多的曝光，例如：

（2）对于重大事件的报道，如何让它入库时就有更高的热度，我们采用的是热词匹配的方式。

即对大型新闻站点的头条，Twitter热点，竞品的头条做监控和扒取，并将这批新闻的关键词维护到热词库并保持更新；每条新闻入库的时候，让新闻的关键词去匹配热词库，匹配度越高，就有越高的初始热度分。

这样处理后，重大事件发生时，Twitter和门户网站的争相报道会导致热词集中化，所有匹配到这些热词的新闻，即报道同样事件的新闻，会获得很高的初始热度分。

9. 新闻摘要提取，支持管理员二次编辑和处理