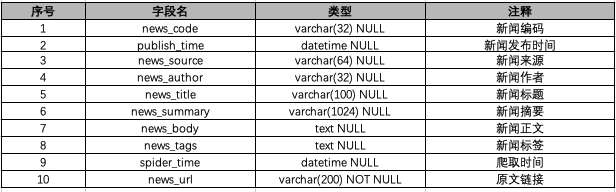
**新闻数据分析模块需求文档**

本模块主要的功能是对新闻爬取数据库的新闻进行分析，了解新闻数据的规律，以便更好的优化新闻源与新闻爬虫。

新闻数据库表结构如下：



分析需求：

1. **数据库整体分析**

分析数据库内新闻进行整体统计分析。一是整体新闻数量的走势，在24小时/1周内的走势，可以折线图的方式来展示；二是分析新闻来源于各个网站所占的比例，可用饼状图的方式展示；三是对数据库表中新闻来源进行分析，分析新闻数据库中来源于这些新闻源的数量，比如按照数量列表展示进行排序，可以以饼状图的方式来展示来源。注意二中爬取网站与新闻源的区别，爬取网站是指新闻抓取的网站。而新闻源是新闻的产生的真实来源。例如：从新浪财经网站爬取的新闻源是证券日报。

1. **网站更新规律分析**

分析单个新闻网更新数量的规律，包括24小时内每个网站的爬取新闻数量，每个频道的新闻爬取数量。以及在每60分钟/30分钟/15分钟内新闻数量的走势，可以柱状图或者折线图的方式来展示。新闻编号中可以解析出新闻来源以及频道信息，具体的规则可以参考《新闻源数据》表。

1. **爬取时间延迟分析**

通过比较publish\_time 和spider\_time两个字断，对新闻源的爬取延时进行分析。统计出60分钟/30分钟/15分钟新闻延迟数据的走势，可以以柱状图或者折线图的方式来表示。

1. **单个爬取网站内新闻源分析**

对单个爬取源网站的新闻源进行分析，分析买个目标爬取网站的新闻来源，比如按照数量列表展示进行排序，可以以饼状图的方式来展示来源。

**备注**

1. 展示效果与设计思路可以参考echart，以最合适的形式展示分析结果。echart网站展示示例链接如下：<https://echarts.baidu.com/examples/>
2. 新闻ID编号规则：



