

基于大数据的优酷网站弹幕的分析与研究

郑海霖 张桂花

四川大学锦城学院计算机与软件学院 四川 成都 611731

【摘要】随着弹幕在视频中的重要性越来越高,弹幕成为了每个大型视频网站的必备品。弹幕和评论相比,它更具有实时性与关键性。所以通过分析研究弹幕来了解观影人的实时情绪也成为了大数据分析的研究热点。主要思路是:通过爬虫获取弹幕信息,使用 HDFS 对数据进行存储,使用 Spark 从 HDFS 上调用数据对弹幕出现的频率进行统计排序并保存。通过研究实现了对观众的实时情绪的了解与视频关键词的分析。

【关键词】爬虫; HDFS; Spark; 弹幕

1 绪论

1.1 研究背景

随着互联网技术的发展视频网站技术日渐成熟,视频网站也越来越多,信息量非常的巨大,人们的工作和生活对视频网站的信息需求也越来越大,所有的视频弹幕网站的重要性愈加明显。并且随着弹幕加入了视频,视频也变得越来越有意思。人们不仅仅会看视频还会注意弹幕的内容,有些弹幕引人深思,有些弹幕充满了愉悦气息,有些弹幕带有一定哲理。而且弹幕不像评论那样是人们通过深思熟虑后的结果,它更具有实时性与关键性。因此弹幕已经成为了视频不可或缺的一部分甚至还能影响视频本身。

1.2 研究意义

随着弹幕在视频中的重要性越来越高,弹幕成为了每个大型视频网站的必备品,而这个原因是有观影人当时行为决定的,所以观影人多,带来的弹幕也是非常之多。有些时候弹幕甚至能够左右视频的质量,一些好的弹幕甚至能够让视频上一个档次。而弹幕是由观影人带来的,所以观影人当时的情绪决定了弹幕的走向。分析弹幕就能够了解观影人当时的想法,还能够了解到人们对这个视频的态度是积极的还是消极的,并且了解视频中的关键信息。多个视频的弹幕可以称之为海量数据,就可以使用大数据进行分析。

2 数据采集

2.1 爬虫原理

通过 Python 来编译。它是一种计算机程序设计语言,同时对编程语言的解释性、编译性、互动性、以及面向对象进行高层次的结合^[1]。网络爬虫技术是指一种按照一定的规律,自动地抓取互联网信息的程序或是脚本^[2]。

2.2 发送请求

首先分析网页获取 URL。打开网站,然后点电视剧播放,点击检查调出浏览器调试窗口。它是指统一资源定位系统,用于定位网页的。

其次,复制任意一条弹幕,然后点击调试窗口按 Control+F 进行搜索。点击该请求的 Headers 按钮,查看请求 URL,观察请求头中的 Referer 和 User-Agent

参数。Headers 包含的是请求头的信息,Referer 是请求头的一部分,表示一个来源,User-Agent 表示用户代理使得服务器能够识别。最后,URL 找到之后便开始编写代码,先从一条数据的抓取,提取,保存,这些都没问题之后再研究批量抓取。

2.3 获取响应内容

提取 json 数据,json 是一种轻量化的数据交换格式,就是为了让浏览器引擎能够识别的数据格式。观察返回的数据会发现,跨域请求都是用的 jsonp,所以需要对返回的数据进行稍微的截取,就是将 URL 中的一部分和最后的部分去掉,只保留中间的 json 数据。

提取弹幕数据,得到 json 之后,就来分析弹幕数据在哪里,在浏览器的调试窗口 Preview 里面查看。Result 字段里面便是弹幕数据,而且数据格式是一个列表,列表中是每个弹幕对线,弹幕对象中的 content 字段就是实际的弹幕内容,直接用 json 把弹幕提取出来。

2.4 解析内容

批量爬取。首先比较分页参数,比较两个请求 URL 的参数,发现同一集的第一次请求与第二次请求的 URL 发现 mat 参数不同,还是依次递增的趋势,这个参数便是寻找的分页参数,将原 URL 中分页参数变为可变参数,由方法传入。然后新建一个批量爬取的方法,循环调用单次爬取方法,每次调用传入页数即可。

比较第一集与第二集的第一个弹幕请求 URL,从而找到不同参数。找到集数参数之后,就可以写一个函数将所有集数参数爬取到。这里需要引入另一个请求头 Cookie (Cookie 是服务器发给客户端的特殊信息,用于服务器记录客户端的状态)。因为 HTTP 协议是无状态协议,服务器不认得你,所以需要用 Cookie 来记录状态。在浏览器调试窗口的 Console 按钮,然后就可以复制出来。

2.5 保存数据

将爬取的数据保存到本地,每条弹幕用逗号作为分字符,用视频名称命名。最后打包成一个文件。

3 数据储存

3.1 HDFS

将保存在本地的数据上传到 HDFS。Hadoop 分布式文件系统，主要运行在通用硬件上 [3]，HDFS 能提供高吞吐量的数据访问，非常适合大规模数据集上的应用。因为获取的数据量非常多，而且需要对数据进行分析 and 处理，如果保存在本地使用 Python 进行处理，效率会非常的低，所以需保存 HDFS 上有利于后续的分析与处理。

3.2 保存数据

使用 xshell 将本地数据传到服务器并解压，直接将要上传的本地文件拖动到对应的服务器位置即可。在 HDFS 上创建一个文件夹，然后将数据放入文件夹内。

4 数据分析和处理

4.1 spark 技术

因为获取的弹幕文件非常多，而且杂乱，如果采用 Python 分析将会浪费大量时间。所以采用 Spark。它是基于内存计算的大数据分布式计算框架，所以有着低延迟的复杂分析，主要为交互式查询和迭代算法设计。相比于 MapReduce，它是基于内存的编程框架，中间结果可存储在内存中，因而它运算速度要高于 MapReduce。因为 Spark 在计算结果的途中可以将数据保存在内存中，所以可以减少读写 HDFS 的次数，增加效率。

4.2 自动化脚本

自动化脚本可以简化操作环节，编写完成以后，下次运行就只用填写相应的参数即可，非常方便节省时间。自动化脚本关键部分演示，只需运行脚本 + 影片名称 + 分析后的影片名称，即可快速进行分析。

4.3 关键分析过程

使用 spark 读取 HDFS 上面的文件。因为数据文件都存在一个大文件中，所以可以用 `val lines = sc.textFile("hdfs://文件名")` 来读取文件。

对读取的文件使用 RDD 算子来处理。RDD 是 Spark 的计算模型，弹性分布式数据集 RDD (Resilient Distributed Dataset)，这种模型可以让用户指定数据储存的位置以及对数据的操作。其中 RDD 有很多算子互相依赖，有助于计算容错。就可以使用如下代码对每条弹幕进行统计计数：

```
val rdd2=lines.flatMap(line => line.
split(“,”)).map(word =>(word,1)).
reduceByKey((a,b)=> a+b)
```

首先 Map 是对集合中每个元素进行操作。flatMap 是对集合中每个元素进行操作然后再扁平化。扁平话意思大概就是先用了一次 map 之后对全部数据再一次 map。Split(“,”) 是对文件中的字符按照逗号进行拆分。map((word =>(word,1)) 将拆分好的字符进行编号，如 (A,1) 的键值对，就是说字符 A 出现了一次。reduceByKey 的作用对象是键值对形式的 RDD，而 reduce 有减少压缩之意，保留一条记录通常有两种结果。一种是只保留我们希望的信息，比如每个键出现的次数。第二种是把值聚合在一起形成列表，这样后续可以对值做进一步的操作。reduceByKey((a,b)=>a+b) 是应用于键值对的数据集时，返回一个新的键值对形成的数据集，其中的每个值是将每个键传递到函数

(a,b)=> a+b 中进行聚合后得到结果如 (“AA”,1)，表示 AA 总共出现了 1 次。如图 1 所示：

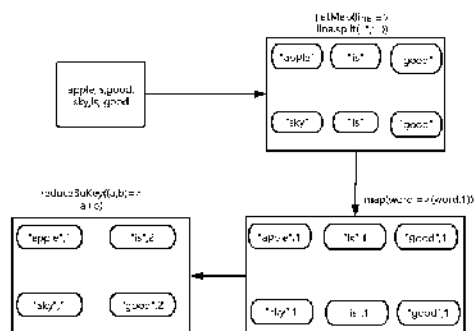


图 1 算法流程

对统计后的数据保存在 HDFS 上面，以读取的文件名来命名。

4.4 分析结果

通过对数据的处理以后，相同弹幕出现的次数由大到小依次呈现眼前。获得如下图结果。可以非常直观的看出高频词与关键词，比如某明星的词频高，就能够告诉我们，本集谁出演了。或者“哈哈”是高频词就能看出本集透露出来了欢乐的气氛。能让我们非常直观的感受当时大部分观影人的心情状态。如下图，脸红共出现了 387 次，结婚共出现了 873 次。有了这样的统计就可以看到非常明确的关键词。比如“完结撒花”就可以看出大部分观影人当时的心情是带有祝福与感慨的，也可以预知到本集会有结婚元素。比如“aws1”可与看出本集会有精彩部分，以及观众的心情是愉快的。比如“前方高能”可看出本集会有剧情的高潮部分，可以看出观众的心情是期待与满意的。比如“津叔”出现了 643 次可以看出本集有津叔出演，并且观众们对于津叔的态度是喜爱的。比如“剧场版”出现了 276 次，可以看出本集观众期待剧场版的出现，并且观众们对于本视频的期待值非常高。比如“名场面预警”可以看出本集中出现了让观众能留下深刻印象的场景，并且观众们对于本视频的某些部分非常认可。通过这样数据处理与分析可以非常直观的了解，观众对此视频的满意度非常高并且心情愉悦，以及会出现的人物和关键事物。结果如图 2 所示：

aws1	3321
官村：在哄了在哄了	2010
要回东西代正确方式	1234
啊喂乱葬岗	1194
完结撒花，感谢陪伴	1003
完结撒花	900
结婚	873
津叔	643
前方高能	600
今日的风儿甚是喧嚣	423
脸红	387
名场面预警	356
剧场版？	276
不愧是你	235
阿坤	190

图 2 结果展示

总结

随着弹幕系统频繁的出现在各大视频网站，看弹幕与发弹幕的用户数剧增。弹幕对视频的影响也在逐步增加，观众与制作者也就越来越关注弹幕，甚至通过弹幕进行互动交流，弹幕的重要性越发显现出来。本文由

此出发,对视频网站弹幕展开研究和分析,首先获取数据,采用大数据技术对数据进行必要的分析和处理,从而定位用户的行为习惯以及用户画像,解决了如何了解观影人的实时情绪与视频关键信息等问题。

【参考文献】

[1] 毕森,杨昱昂.基于python的网络爬虫技术研究[J].数字通信世界,2019(12):107-108.

[2] 顾勤.网络爬虫技术原理及其应用研究[J].信息与电脑(理论版),2021,33(04):174-176.

[3] 刘杨.实践教学强化背景下法学实验室建设的探索与研究[J].科技资讯,2015,13(08):198.
vhj29499@163.com