**《数据科学与数据分析》课程作业**

**专业：** 信息管理学院 **学号：** 211820073 **姓名：** 胡涂

**专业：** 信息管理学院 **学号：** 211840182 **姓名：** 邓雨茵

**专业：** 信息管理学院 **学号：** 211820274 **姓名：** 郑宇

# 在线文章热度的冷启动识别

**摘要：**本文

**关键词：**热度

## 介绍

### 研究背景

在线文章，例如各类社交平台以及新闻平台上的博客以及文章，已经成为现今最重要的信息来源。在线文章的热度，即受欢迎程度，可以用文章的转发量来衡量。

现有的研究中，多数对在线新闻文章的预测，是采用文章时间过程特征以及文本内容特征融合，混合预测高热度文章和与低热度文章。如将时间过程与内容特征进行深度融合对微信公众号文章的浏览量进行预测，在不平衡测试集中对热度分类的预测准确率达到了0.965。

这样的预测方法在工业应用中的价值很高，同时预测准确率也较高。然而对于内容创作者而言会产生一些问题，如高热度文章具有极大的偶然性，对于创作者来说很难从中得到启示，并将经验应用到自己文章之中；很难在无法获得时间过程特征的情况下保持较高的预测准确率。

这样的问题与解决文章的“冷启动问题”的描述相符。文章的“冷启动问题”，即在发表前预测文章热度的方法，现在仍旧在起步阶段。过往研究认为，文章的发布平台是文章热度的决定性因素。如果内容创作者想要调整文章的结构等特征，这将是没有帮助的。同时对冷启动问题，也有研究基于文章标题特征，进行标题的特征建模，从而对新闻热度进行回归。在他们的研究中，对文章热度的回归模型MAE值达到了0.67，这是在对响应变量进行对数转换得出的。本文认为，这样的误差是不能接受的。

同时，过往所有的研究都将低热度和高热度合并当作研究重点。

### 任务介绍

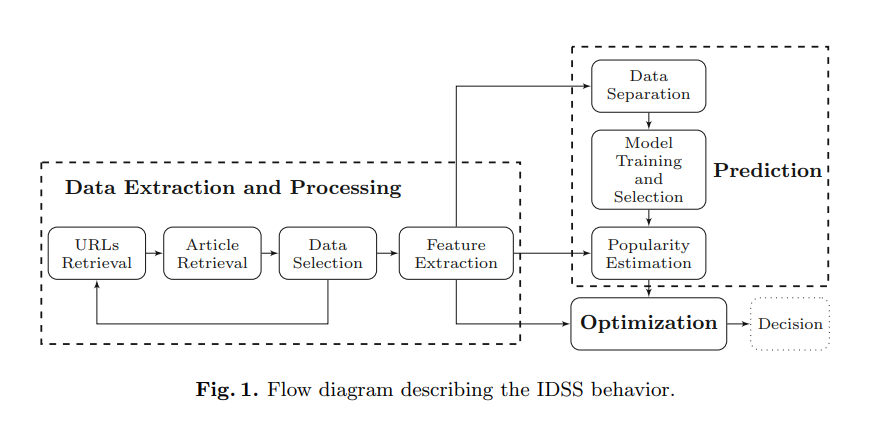
前面提到，高热度文章的偶然性与热点的强关联性很难让内容创作者得到有用的启示。

而极低热度文章，它们或许具有一些另创作者感兴趣的内容特征，能够使文章避免低热度，使得文章能够更关注于内容创作。

为了能够更好地向内容创作者提出建议，便于文章摆脱低热度，能够让内容获取者更注重文章内容。我们采取手工制作文章特征的方式，基于文章文本特征识别低热度文章与正常文章的任务，同时对文本特征进行相应解释，以便向内容创作者提出修改建议。

为了解决这一问题，我们将采用数据获取-特征构建-模型拟合-模型解释的流程。

（这边放一整张图片，决策支持系统的）



## 数据获取与特征构建

数据字典是对数据集中的各个特征进行解释的文档，它可以帮助我们理解数据的含义和来源。本文根据不同的特征类型，将数据字典分为以下几个部分：

（一）数据集获取

本研究使用的数据集来源于https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity，该数据集包含了两年内由Mashable发布的文章的各种特征和社交媒体分享数。原数据集共有39797条记录和61个特征，我们根据研究目的对特征进行了筛选，并且添加了一些新的特征。

（二）特征的添加

我们利用Python网络爬虫技术根据数据集中提供的网址，爬取了文章的标题和正文内容，并且将其转化为文本格式。同时，考虑到新闻文章的热度与其所报道的事物的热度密切相关，我们还爬取了文章所属主题在文章发布前一周内的谷歌趋势指数。（三）特征的选择

### 1、时间特征

文章发布时间会影响文章的新颖性和时效性，从而影响读者的兴趣和分享意愿。一般来说，对于同一个热点事物，越早发布的文章越能抓住读者的注意力，越容易被分享。

|  |  |  |
| --- | --- | --- |
| 特征 | 含义 | 属性 |
| timedelta | 距离收集时间的间隔，以天为单位。 | Number |
| Date | 集数据的日期，格式为年-月-日。 | Date |
| dayRatio | 文章发表的前一天相较于前两天的热度增长率 | Number(0~1) |
| threeDayRatio | 文章发表的前三天相较于前六到三天的热度的增长率 | Number(0~1) |
| weekRatio | 文章发表的前一周相较于前两周的热度的增长率 | Number(0~1) |
| twoWeekRatio | 文章发表的前两周相较于前四到二的热度的增长率 | Number(0~1) |

文章的发布时间可能与文章的主题和内容相关，从而影响读者的偏好和分享动机例如，一些文章可能与特定的节假日或事件有关，如果在适当的时间发布，可能会引起更多的共鸣和分享。

|  |  |  |
| --- | --- | --- |
| 特征 | 含义 | 属性 |
| weekday\_is\_monday | 是否是星期一，是为1，否为0(下同)。 | Factor(0,1) |
| weekday\_is\_tuesday | 是否是星期二 | Factor(0,1) |
| weekday\_is\_wednesday | 是否是星期三 | Factor(0,1) |
| weekday\_is\_thursday | 是否是星期四 | Factor(0,1) |
| weekday\_is\_friday | 是否是星期五 | Factor(0,1) |
| weekday\_is\_saturday | 是否是星期六 | Factor(0,1) |
| weekday\_is\_sunday | 是否是星期日 | Factor(0,1) |
| is\_weekend | 是否是周末 | Factor(0,1) |
| isHoliday | 是否是节假日 | Factor(0,1) |
| HolidayName | 如果是节假日，节假日名称（Not Holiday表示不是节假日）。 | Factor |
| HolidayDay | 如果是节假日，是节假日的第几天（从1开始计数）。 | Number |
| HolidayDaysLeft | 如果是节假日，距离节假日结束还有几天（最后一天算1）。 | Number |

### 2、栏目特征

栏目可能影响文章的主题和内容，从而影响读者的兴趣和偏好，以及分享意愿。一般来说，不同的栏目可能吸引不同的读者群体，有些栏目可能更受欢迎或更有争议，从而更容易被分享。

[1]李泽月. 基于新浪新闻数据分析的新闻热度预测方法研究[D].中国地质大学(北京),2019.DOI:10.27493/d.cnki.gzdzy.2019.000950.

|  |  |  |
| --- | --- | --- |
| 特征 | 含义 | 属性 |
| data\_channel\_is\_lifestyle | 是否属于生活方式栏目，是为1，否为0 | Factor(0,1) |
| data\_channel\_is\_entertainment | 是否属于娱乐栏目 | Factor(0,1) |
| data\_channel\_is\_bus | 是否属于商业栏目 | Factor(0,1) |
| data\_channel\_is\_socmed | 是否属于社交媒体栏目 | Factor(0,1) |
| data\_channel\_is\_tech | 是否属于科技栏目 | Factor(0,1) |
| data\_channel\_is\_world | 是否属于世界栏目 | Factor(0,1) |

### 3、标题特征

标题字数、比较级与最高级词数、强化词与弱化词数、标点符号个数、不同词性的词数、标题主观性和标题情感极性会影响文章的吸引力和刺激性，从而影响读者的点击率和分享意愿。一般来说，文章的吸引力和刺激性越强，文章越能激发读者的好奇心和情感，越容易被分享，但也可能被认为是哗众取宠的表现而被忽视。

|  |  |  |
| --- | --- | --- |
| 特征 | 含义 | 属性 |
| n\_tokens\_title | 标题字数，以空格分隔的单词数 | Number |
| Comparatives Count | 标题比较级个数 | Number |
| Superlatives Count | 标题最高级个数 | Number |
| Count Intensifiers | 标题强化词个数 | Number |
| Count Downtoners | 标题弱化词个数 | Number |
| noun\_count | 标题名词个数 | Number |
| verb\_count | 标题动词个数 | Number |
| adverb\_count | 标题副词个数 | Number |
| punc\_count | 标题标点符号个数 | Number |
| title\_subjectivity | 标题主观性，越高表示标题越带有个人情感或观点 | Number(0~1) |
| title\_sentiment\_polarity | 标题情感极性，越高表示标题越积极或正面。 | Number(0~1) |

标题的可读性、简洁性、准确性和新奇性对于吸引读者和促进分享非常重要。本研究探讨了四个影响标题可理解性的特征：可读性指数、语法分析树高度、歧义和新奇程度。可读性指数反映了标题的阅读难度，我们使用 Flesch-Kincaid Grade Level 公式来计算。语法分析树高度反映了标题的结构复杂度，我们使用Python中的NLTK中的句法分析器来生成和分析。歧义反映了标题的表达清晰度，我们使用词语含义数的乘积来估计。新奇程度我们使用当前标题tf-idf的和在它之前的标题的tf-idf余弦相似度最大值得到。我们发现，这些特征越低，标题越新奇，容易阅读和分享；但是如果可读性、简洁性、准确性过低，标题可能会缺乏复杂性和深度，从而降低其价值和信任度。

|  |  |  |
| --- | --- | --- |
| 特征 | 含义 | 属性 |
| **Flesch Kincaid Grade of Title** | 标题可读性指数 | Number |
| **SyntaxTree Height** | 标题语法分析树高度 | Number |
| All Possible Meanings | 标题所有可能意思（歧义） | Number |
| novel of title | 新奇程度 | Number(0~1) |

### 4、正文特征

文章的内容丰富程度是指文章能够提供多少有价值和有意义的信息给读者，从而满足读者的需求和期望。文章的正文字数、不同字数占比、非停用词占比和非停用词不同字数占比反映了文章的长度和信息量。外链和自链可以增加文章的信息来源和参考资料，扩展读者的知识面和视野，从而提高文章的内容丰富程度。图片和视频可以增加文章的视觉效果和趣味性，丰富读者的感官体验和情感反应，从而提高文章的内容丰富程度。文章中名词，形容词和动词的占比反映了文章的语言风格。名词占比较高的文章可能更客观和专业，形容词占比较高的文章可能更主观和描述性，动词占比较高的文章可能更动态和有说服力。

|  |  |  |
| --- | --- | --- |
| 特征 | 含义 | 属性 |
| n\_tokens\_content | 正文字数 | Number |
| n\_unique\_tokens | 正文不重复的单词数 | Number |
| n\_non\_stop\_words | 正文非停用词数 | Number |
| n\_non\_stop\_words | 正文非停用不重复的单词数 | Number |
| num\_hrefs | 外链个数，指向其他网站的超链接数。 | Number |
| num\_self\_hrefs | 自链个数，指向本网站的超链接数。 | Number |
| num\_imgs | 图片个数 | Number |
| num\_videos | 视频个数 | Number(0~1) |
| NWordRatio | 名词占比 | Number(0~1) |
| JWordRatio | 形容词占比 | Number(0~1) |
| VWordRatio | 动词占比 | Number(0~1) |

文章的主观程度和情感极性是指文章在多大程度上反映作者的个人看法和态度，以及作者对于文章主题的正面或负面的评价。主观程度越高，文章越倾向于表达作者的个人观点和情感，主观程度越低，文章越倾向于表达客观事实和数据。情感极性越高，文章越倾向于表达正面的情感，情感极性越低，文章越倾向于表达负面的情感。文章的主观程度和情感极性可以影响读者对文章的认同和共鸣，从而影响文章的传播效果。

|  |  |  |
| --- | --- | --- |
| 特征 | 含义 | 属性 |
| global\_subjectivity | 文章的主观程度(越大越主观) | Number(0~1) |
| globalsentimentpolarity | 文章的情感极性(正代表积极，负代表消极，绝对值越大情感越强烈) | Number(-1~1) |
| globalratepositive\_words | 积极词的占比 | Number(0~1) |
| globalratenegative\_words | 消极词的占比 | Number(0~1) |
| ratepositivewords | 非中性词积极词的占比 | Number(0~1) |

文章的可读性是指文章有多容易被读者理解，我们使用FleschReadingEase公式和文章所用词在牛津8000词的词的占比来衡量这一指标。文章越简单和清晰，可读性越低，文章越复杂和晦涩。

|  |  |  |
| --- | --- | --- |
| 特征 | 含义 | 属性 |
| ContentFleschReadingEase | 可读性 | Number |
| wordRatioIn8000 | 在牛津8000词的词的占比 | Number |

### 5、预测目标

shares： 文章被分享次数。

## 模型拟合与优化过程

## 总结

附录

时间特征

timedelta： 距离收集时间的间隔，以天为单位。

Date： 收集数据的日期，格式为年-月-日。

weekday\_is\_monday： 是否是星期一，是为1，否为0。

weekday\_is\_tuesday： 是否是星期二，是为1，否为0。

weekday\_is\_wednesday： 是否是星期三，是为1，否为0。

weekday\_is\_thursday： 是否是星期四，是为1，否为0。

weekday\_is\_friday： 是否是星期五，是为1，否为0。

weekday\_is\_saturday： 是否是星期六，是为1，否为0。

weekday\_is\_sunday： 是否是星期日，是为1，否为0。

is\_weekend： 是否是周末，是为1，否为0。

isHoliday： 是否是节假日，是为1，否为0。

HolidayName： 如果是节假日，节假日名称（Not Holiday表示不是节假日）。

HolidayDay： 如果是节假日，是节假日的第几天（从1开始计数）。

HolidayDaysLeft： 如果是节假日，距离节假日结束还有几天（最后一天算1）。

dayRatio： 文章发表的前一天相较于前两天的热度增长率增长

threeDayRatio： 文章发表的前三天相较于前六到三天的热度增长率增长

weekRatio：文章发表的前一周相较于前两周的热度增长率增长

twoWeekRatio：文章发表的前两周相较于前四到二的热度增长率增长

栏目特征

data\_channel\_is\_lifestyle： 是否属于生活方式栏目，是为1，否为0。

data\_channel\_is\_entertainment： 是否属于娱乐栏目，是为1，否为0。

data\_channel\_is\_bus： 是否属于商业栏目，是为1，否为0。

data\_channel\_is\_socmed： 是否属于社交媒体栏目，是为1，否为0。

data\_channel\_is\_tech： 是否属于科技栏目，是为1，否为0。

data\_channel\_is\_world： 是否属于世界栏目，是为1，否为0。