**《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

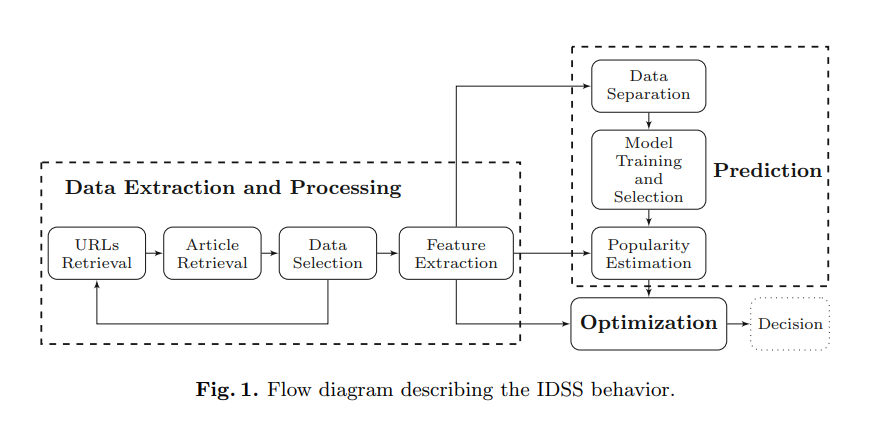
前面提到，高热度文章的偶然性与热点的强关联性很难让内容创作者得到有用的启示。

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：

时间特征

文章发布时间可能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，越早发布的文章越能抓住读者的注意力，越容易被分享。

文章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。

文章发布时间可能与文章的质量和可信度有关，从而影响读者的评价和分享行为。例如，一些文章可能经过了更多的审稿和修改，因此发布时间较长，但可能会提高文章的科学性和准确性，从而增加读者的信任和分享意向。

综上所述，文章发布时间与文章分享量之间可能存在正向或负向的关系，具体取决于文章的类型、目标和受众。因此，作者在选择合适的发布时间时，应该考虑到这些因素，并根据自己的研究目标和期望进行权衡和决策。

发表时间

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）。

热度变化（使用绝对热度的效果应该会好一些，但是Google指数所给的是相对的热度，因此只能采取这种方法）

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

描述性特征

正文字数、外链个数、图片个数和视频个数可能影响文章的内容丰富度和多样性，从而影响读者的阅读体验和分享意愿。一般来说，这些特征越多，文章越能吸引读者的注意力，越容易被分享。

正文不同字数占比、正文非停用词占比和正文非停用词不同字数占比可能影响文章的语言质量和创新性，从而影响读者的认知评价和分享动机。一般来说，这些特征越高，文章越能体现作者的独特见解和表达能力，越容易被分享。

平均单词长度可能影响文章的可读性和专业性，从而影响读者的理解难度和分享行为。一般来说，这个特征越低，文章越容易阅读，越容易被分享；但是如果过低，文章可能会失去一些专业术语和精确度，从而降低文章的权威性和信任度。

综上所述，描述性特征里面的特征与文章分享量之间可能存在正向或负向的关系，具体取决于文章的类型、目标和受众。因此，作者在选择合适的描述性特征时，应该考虑到这些因素，并根据自己的研究目标和期望进行权衡和决策。

n\_tokens\_title: 标题字数，以空格分隔的单词数。

n\_tokens\_content: 正文字数，以空格分隔的单词数。

n\_unique\_tokens: 正文不同字数占比，即正文中不重复的单词数除以正文总单词数。

n\_non\_stop\_words: 正文非停用词占比，即正文中除去常用词（如a, the, and等）的单词数除以正文总单词数。

n\_non\_stop\_unique\_tokens: 正文非停用词不同字数占比，即正文中除去常用词的不重复单词数除以正文总单词数。

num\_hrefs: 外链个数，即正文中指向其他网站的超链接数。

num\_self\_hrefs: 自链个数，即正文中指向本网站的超链接数。

num\_imgs: 图片个数，即正文中包含的图片数。

num\_videos: 视频个数，即正文中包含的视频数。

average\_token\_length: 平均单词长度，即正文中所有单词的字符数之和除以正文总单词数。

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。

栏目可能影响文章的数据共享和可重用性，从而影响读者的认知评价和分享动机。一般来说，一些栏目可能涉及更多的数据和方法，如果能够有效地共享和重用这些数据，可能会提高文章的科学性和价值，从而增加读者的信任和分享意向。

综上所述，栏目与分享量之间可能存在正向或负向的关系，具体取决于文章的类型、目标和受众。因此，作者在选择合适的栏目时，应该考虑到这些因素，并根据自己的研究目标和期望进行权衡和决策。

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。

标题特征

标题字数、比较级个数、最高级个数、强化词个数、标点符号个数可能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，这些特征越多，文章越能激发读者的好奇心和情感，越容易被分享。¹

标题可读性指数、语法分析树高度、所有可能意思可能影响文章的可理解性和歧义性，从而影响读者的阅读难度和分享动机。一般来说，这些特征越低，文章越容易阅读，越容易被分享；但是如果过低，文章可能会失去一些复杂性和深度，从而降低文章的价值和信任度。

标题主观性、标题情感极性、标题新颖性可能影响文章的观点和内容，从而影响读者的认同感和分享行为。一般来说，这些特征越高，文章越能体现作者的态度和创造力，越容易被分享。

综上所述，标题提炼出的特征对分享量的可能的影响取决于文章的类型、目标和受众。因此，作者在选择合适的标题特征时，应该考虑到这些因素，并根据自己的研究目标和期望进行权衡和决策。

topicNo: LDA主题所属类别编号（从0开始计数）。

Comparatives Count: 标题比较级个数（如better, faster等）。

Superlatives Count: 标题最高级个数（如best, fastest等）。

Count Intensifiers: 标题强化词个数（如very, really等）。

Count Downtoners: 标题弱化词个数（如kind of, sort of等）。

Flesch Kincaid Grade of Title: 标题可读性指数（简洁性），越低越容易阅读。计算公式：(0.39 \* 平均句子长度) + (11.8 \* 平均音节数) 15.59。其中平均句子长度指标题总单词数除以标题总句子数；平均音节数指标题总音节数除以标题总单词数。音节数指一个单词中元音字母出现的次数（连续元音只算一次），通常与音节相同或接近。例如：cat有一个音节数；education有四个音节数；idea有三个音节数；beautiful有三个音节数；cooperate有三个音节数；queue有一个音节数等等。

SyntaxTree Height: 标题语法分析树高度（简洁性），越低越容易阅读。语法分析树指将一个句子按照语法规则分解成不同的成分，并用树状结构表示出来。

All Possible Meanings: 标题所有可能意思（歧义），越多越难理解。例如：I saw a man on a hill with a telescope. 可能有以下几种意思：

我在山上看到了一个拿着望远镜的男人。（我在山上；男人拿着望远镜）

我用望远镜看到了一个在山上的男人。（我拿着望远镜；男人在山上）

我看到了一个在山上用望远镜观察的男人。（男人在山上；男人拿着望远镜）

我看到了一个站在带有望远镜的山上的男人。（山上有望远镜；男人在山上）

novel of title：表示当前标题和在它之前的tf-idf余弦相似度最大值（新奇性），越小越新颖。tf-idf指一种衡量一个单词对于一个文档或者一个语料库重要程度的方法。

noun\_count: 标题名词个数

verb\_count: 标题动词个数

adverb\_count: 标题副词个数

punc\_count: 标题标点符号个数（如! , ?等）。

title\_subjectivity: 标题主观性，越高表示标题越带有个人情感或观点。

title\_sentiment\_polarity: 标题情感极性，越高表示标题越积极或正面。

正文特征

正文主观性、正文情感极性、正文积极词占比、正文消极词占比、正文积极词比例、正文消极词比例、正文积极词平均情感极性、正文消极词平均情感极性可能影响文章的情感表达和传播，从而影响读者的情感共鸣和分享意愿。一般来说，这些特征越高，文章越能激发读者的正面情感，越容易被分享。

正文不同字数占比、正文非停用词占比、正文非停用词不同字数占比可能影响文章的语言质量和创新性，从而影响读者的认知评价和分享动机。一般来说，这些特征越高，文章越能体现作者的独特见解和表达能力，越容易被分享。

好的，我会尽力帮你论述一下。根据网上的一些资料¹²³，新闻文章分享量的影响因素主要有以下几个方面：

- \*\*内容可读性\*\*：指文章的语言是否通顺，逻辑是否清晰，结构是否合理，是否符合受众的阅读习惯和水平。内容可读性高的文章能够吸引受众的注意力，提高阅读兴趣和满意度，从而增加分享的可能性。

- \*\*在8000词的词的占比\*\*：指文章中使用的词汇是否丰富，是否有新颖和独特的表达方式，是否能够准确和生动地传达信息。在8000词的词的占比高的文章能够展示作者的文采和专业性，增加文章的吸引力和说服力，从而提高分享的意愿。

- \*\*名词占比\*\*：指文章中使用的名词是否具有代表性，是否能够突出文章的主题和重点，是否能够引发受众的共鸣和关注。名词占比高的文章能够提高文章的信息量和价值，增加受众的认同感和信任感，从而促进分享的行为。

- \*\*形容词占比\*\*：指文章中使用的形容词是否恰当，是否能够增强文章的描述性和表现力，是否能够激发受众的情感和想象。形容词占比高的文章能够提高文章的美感和感染力，增加受众的喜爱度和赞赏度，从而激发分享的动机。

- \*\*动词占比\*\*：指文章中使用的动词是否精准，是否能够展示文章的动态和变化，是否能够引导受众的行动和参与。动词占比高的文章能够提高文章的生动度和活力，增加受众的兴趣度和参与度，从而推动分享的过程。

- \*\*文章主观程度\*\*：指文章中作者对于事实或观点所持有的态度或立场，以及所表达出来的情感或评价。文章主观程度高的文章能够展示作者的个性和风格，增加文章的独特性和争议性，从而引发受众的讨论和互动，促进分享的交流。

- \*\*文章情感极性\*\*：指文章中所传达出来的正面或负面情感倾向，以及所引起的受众情感反应。文章情感极性强烈的文章能够影响受众的心理状态和情绪变化，增加文章

ContentFleschReadingEase:可读性

wordRatioIn8000: 在8000词的词的占比

NWordRatio: 名词占比

JWordRatio: 形容词占比

VWordRatio: 动词占比

global\_subjectivity :文章的主观程度

globalsentimentpolarity: 文章的情感极性

globalratepositive\_words: 积极词的占比

globalratenegative\_words: 消极词的占比

ratepositivewords 非中性词积极词的占比

shares: 文章被分享次数（目标变量）。

## 模型拟合与优化过程

## 总结