

课程设计报告

**《信息系统集成与开发》**

姓 名： 高政

学 号： 19851242

专 业： 计算机科学与技术

系 别： 计算机与信息技术

指导教师： 王志海

**2022年6月23日**

# B站优质UP主分析报告

此次分析报告我将使用和鲸社区中的数据进行分析，通过大量数据分析用户对于B站中视频分类的喜爱，分析软件使用Weka进行分析。

首先我们先要对数据进行清洗，剔除一些空数据、重复以及无效数据；后我们将计算每个UP主的IFL模型数据，之后我们将进行对数据的分析，将会使用K-Means算法，中文名称为K均值聚类算法，是一种迭代求解的聚类分析算法；最后，当模型训练完成后，我们会得到k个分组，我们将每个分组在一起的数据分为一蔟，我们对每一蔟数据进行比较，分析其中的不同

## 1 数据分析网站简介

Weka的全名是怀卡托智能分析环境，是一款免费的，非商业化的，基于Java环境下开源的机器学习以及数据挖掘软件，Weka的主要开发者来自新西兰[1]。

Weka是用于数据挖掘认为的机器学习算法的集合。它包含用于数据准备、分类、回归、聚类、关联规则挖掘和可视化的工具。Weka只在新西兰的岛屿上发现，是一种不会飞的鸟，具有好奇的天性。这个名字的发音是这样的，鸟儿听起来像这样。Weka是根据GNU通用公共许可证发布的开源软件。

### 1.1 Weka及其团队简介

机器学习方法（ML）的发明和应用是计算机科学中一个激动人心且可能影响深远的发展。这些使计算机程序能够自动分析大量数据，并确定哪些信息最相关。这些具体化的信息可以用来自动做出预测，或帮助人们更快、更准确地做出决策。

Weka的团队已经将几种标准的ML技术集成到一个名为Weka的软件“工作台”中，用于Waikato环境的知识分析。有了它，特定领域的专家能够使用ML从太大而无法手动分析的数据库中获取有用的知识。Weka的用户是ML研究人员和工业科学家，但它也广泛用于教学。最近，此团队还致力于MOA，一种用于挖掘数据流的环境。

Weka是用于数据挖掘任务的机器学习算法的集合。它包含用于数据准备、分类、回归、聚类、关联规则挖掘和可视化的工具。同时Weka是根据GNU通用公共许可证发布的开源软件，并且支持深度学习[2]。

### 1.2 项目目标

Weka项目目标共有四个：

(1). 使ML技术普遍可用；

(2). 将其应用于对新西兰工业至关重要的实际问题；

(3). 开发新的机器学习算法并向世界推广；

(4). 为该领域的理论框架做出贡献。

## 2 软件系统功能

Weka系统汇集了前沿的机器学习算法和数据预处理工具，以便用户能够快速灵活地将已有的成熟处理方法应用于新的数据集。

### 2.1 Weka基本介绍

接下来是对Weka的简单介绍

#### 2.1.1 Weka的历史

1992年末，新西兰怀卡托大学计算机科学系Ian Written博士申请基金。

1993年获新西兰政府资助，并于同年开发出接口和基础架构。

1994年发布了第一个Weka的内部版本。

1996年，第一个公开版本Weka2.1发布。Weka的早期版本主要采用C语言编写。

1997年，团队决定使用Java重新改写。Java允许用一个统一的接口来学习方案和方法的预处理和后处理，决定使用Java来代替C++或其他面向对象的语言，是因为Java编写的程序可以运行在绝大部分计算机上，而无须重新编译，更不需要修改源代码。Weka已经测试过的平台包括Linux, Windows和Macintosh操作系统，甚至PDA。最后的可执行程序复制过来即可运行，完全绿色，不需求复杂安装。当然，Java也有其缺点，最大的问题是它在速度上有缺陷，执行一个Java程序比相应的C语言程序马上好几倍。

1999年，发布Java的Weka 3版本。

2013年，Weka的最新版本是3.7.8。

#### 2.1.2 Weka系统简介

Weka作为一个公开的数据挖掘工作平台，集合了大量能承担数据挖掘任务的机器学习算法，包括对数据进行预处理，分类，回归、聚类、关联规则以及在新的交互式界面上的可视化。

如果想自己实现数据挖掘算法的话，可以参考Weka的接口文档。在Weka中集成自己的算法甚至借鉴它的方法自己实现可视化工具并不是件很困难的事情。

2005年8月，在第11届ACM SIGKDD国际会议上，怀卡托大学的Weka小组荣获了数据挖掘和知识探索领域的最高服务奖，Weka系统得到了广泛的认可，被誉为数据挖掘和机器学习历史上的里程碑，是现今最完备的数据挖掘工具之一（已有11年的发展历史）。Weka的每月下载次数已超过万次。

2014年3月起，新西兰怀卡托大学将推出Weka免费网课，课程分为初级和高级两个部分，每个部分时长5周。初级课程将于2014年3月3日开课，高级课程于2014年4月下旬开课。课程具体内容参见怀卡托大学网站Weka MOOC。课程在优酷网站也有专辑。

#### 2.1.3 输入数据

输入的数据有两种：

(1). 通过以ARFF格式为代表的文件进行输入

(2). 直接读取数据库表

### 2.2 Weka界面简介

Weka系统汇集了最前沿的机器学习算法和数据预处理工具，以便用户能够快速灵活地将已有的处理方法应用于新的数据集。

Weka主界面为Weka GUI选择器，它通过右边的四个按钮提供四种主要的应用程序供用户选择，如图1.1。

#### 2.5.1 Explorer（探索者）

系统提供的最容易使用的图像用户接口。通过选择菜单和填写表单，可以调用Weka的所有功能。

虽然探索者界面使用很方便，但它也存在一个缺陷，要求它将所需数据全部一次读进内存，一旦用户打开某个数据集，就会读取全部数据。因此，这种批量方式仅适合处理中小规模的问题。而**知识流**刚好能够弥补这一缺陷。

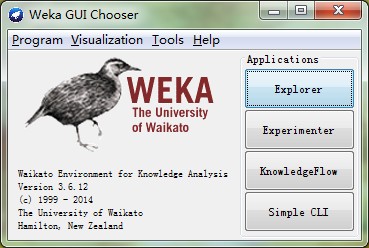


图1.1 Weka软件首页

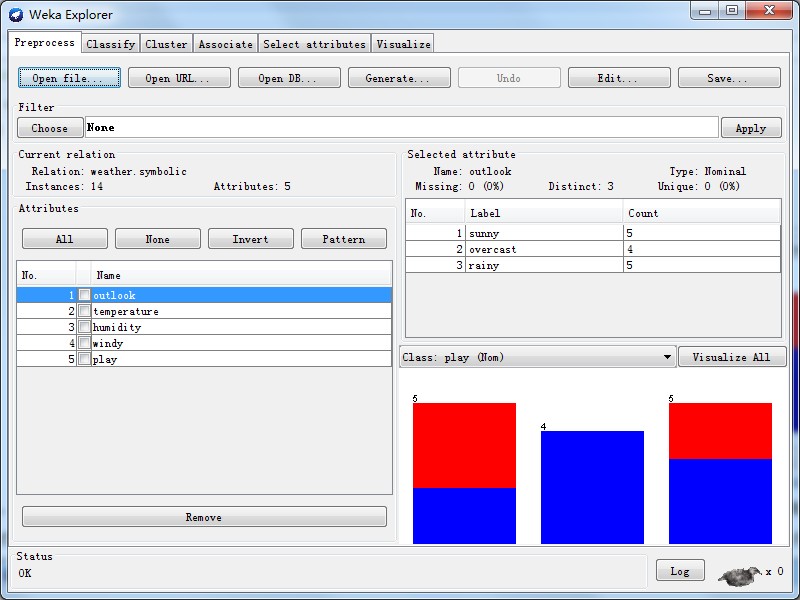


图1.2 Explorer页面

#### 2.5.2 Knowledge Flow（知识流）

使用增量方式的算法来处理大型数据集，用户可以定制处理数据流的方式和顺序。按照一定顺序将代表数据源、预处理工具、学习算法、评估手段和可视化模块的各构件组合在一起，形成数据流。

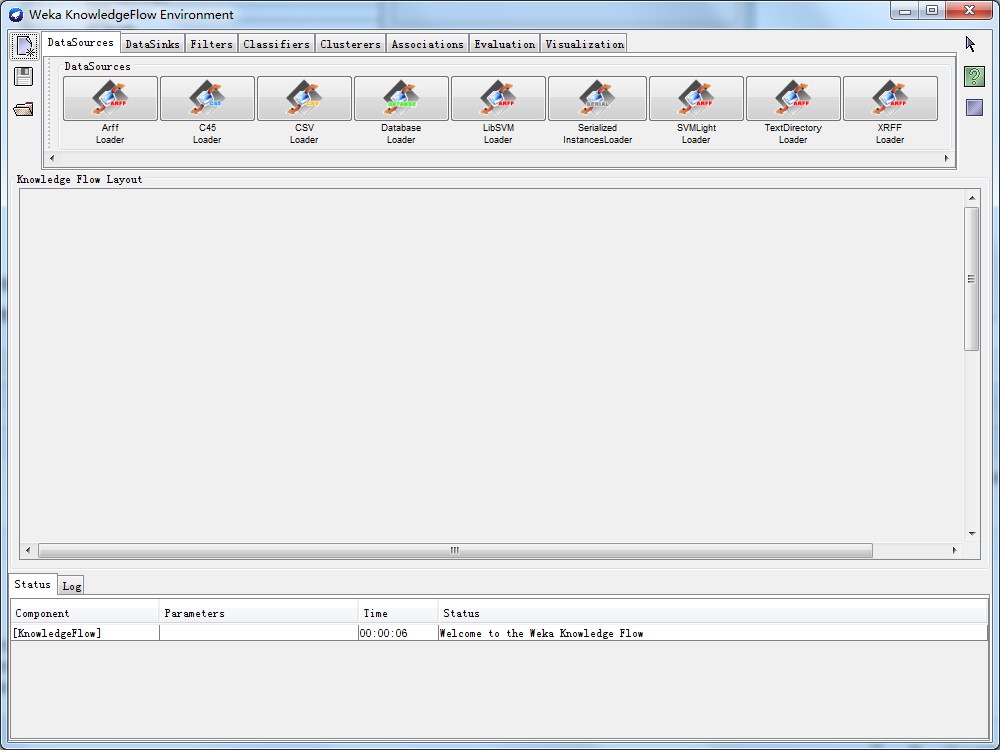


图1.3 Knowledge Flow页面

#### 2.5.3 Experimenter（实验者）

用于帮助用户解答实际应用分类和回归技术中遇到的一个基本问题-----对于一个已知问题，哪种方法及参数值能够取得最佳效果？尽管**探索者界面**也能通过交互完成这样的功能，但通过**实验者界面**，用户可以让处理过程实现自动化。实验者界面更加容易使用不同参数去设置分类器和过滤器，使之运行在一组数据集中，收集性能统计数据，实现重要的测试实验。

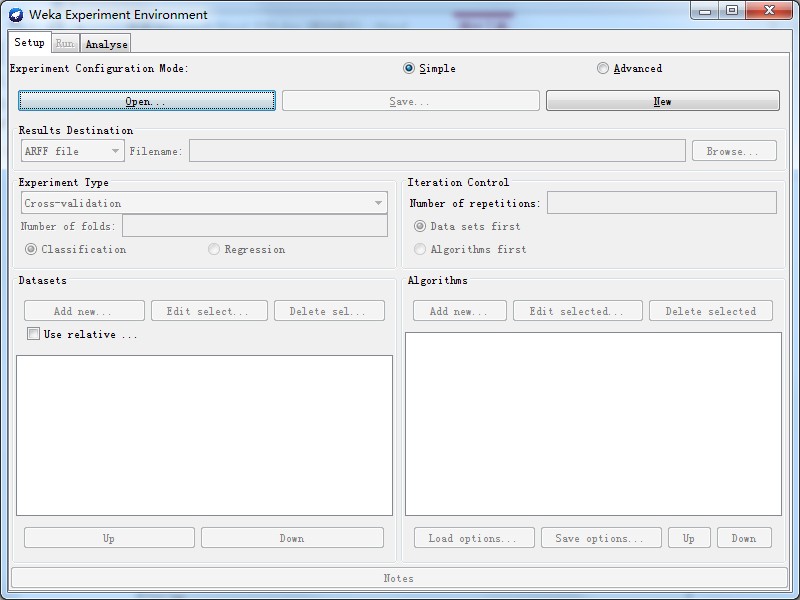


图1.4 Experimenter页面

#### 2.5.4 Simple CLI（简单命令行）

这个界面是为不提供自己的命令行界面的操作系统提供的，该简单命令行界面用户和用户进行交互，可以直接执行Weka命令[3]。

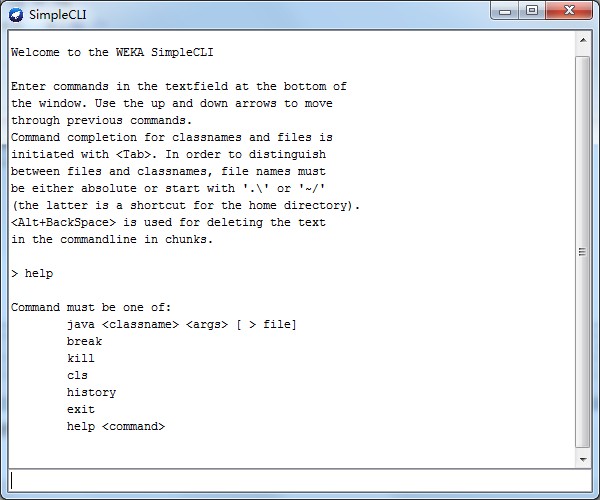


图1.5 Simple CLI页面

### 2.3 Weka处理方法

包括处理标准数据挖掘问题的所有方法：回归、分类、聚类、关联规则和属性选择。

#### 2.3.1 ****分类和回归****

分类和回归是预测的两种形式，分类预测输出目标是离散值，回归预测的输出目标是连续值。

分类的步骤：第一步，分析训练集的特点并构建分类模型；第二步，使用构建好的分类模型对预测集进行分类，评估分类模型的分类准确度等指标。

（1）决策树分类

（2）贝叶斯分类

（3）k-最近邻分类

（4）神经网络分类

#### 2.3.2 ****聚类分析****

聚类为无监督学习方法。主要方法有：

（1）k均值算法

（2）EM（期望最大化）算法

（3）DBSCAN算法

#### 2.3.3 ****关联分析****

用于发现隐藏在大型数据集中有意义的联系，这种联系可以用关联规则进行表示。

主要方法有：

（1）Apriori算法

（2）FP-Growth（频繁模式增长）算法

### 2.3 机器学习概念

机器学习是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。

它是人工智能核心，是使计算机具有智能的根本途径。

以下我们将从机器学习的定义、分类、测试方法、、来分别介绍机器学习的最基本概念。

#### 2.3.1 机器学习的定义

机器学习是一门多学科交叉专业，涵盖概率论知识，统计学知识，近似理论知识和复杂算法知识，使用计算机作为工具并致力于真实实时的模拟人类学习方式，并将现有内容进行知识结构划分来有效提高学习效率。

机器学习有下面几种定义：

（1）机器学习是一门人工智能的科学，该领域的主要研究对象是人工智能，特别是如何在经验学习中改善具体算法的性能。

（2）机器学习是对能通过经验自动改进的计算机算法的研究。

（3）机器学习是用数据或以往的经验，以此优化计算机程序的性能标准。

#### 2.3.2 机器学习的分类

机器学习的分类有很多种，本次仅讨论基于学习方式的分类，有三种。

（1）有监督学习（预测学习），目标是在给定一系列输入/输出实例所构成的数据集的条件下，学习输入x到输出y的映射关系。这里的数据集称为训练集，实例的个数称为训练样本数。（从分类角度看，即训练集中各组数据的类别已知）可以对所观察到的值与预测值进行比较，得到明确的误差值。

（2）无监督学习（描述学习），目标是在给定一系列输入实例构成的数据集的条件下，发现数据中的有趣模式。因为我们不知道需要寻找什么样的模式，也没有明显的误差度量可供使用。

（3）强化学习（增强学习）：以环境反馈（奖/惩信号）作为输入，以统计和动态规划技术为指导的一种学习方法。

#### 2.3.3 测试方法

机器学习的测试方法大致分为4种类型：

（1）将全部数据进行训练，从而建立一个模型，然后使用这个模型对全部数据进行训练测试。

（2）将全部数据分为两部分，一部分用来训练模型，一部分用来测试模型。

（3）分别提供两个数据集，一个是训练集合，一个是测试集合。

（4）k重交叉验证，十重交叉验证或k重交叉验证。把一个数据集合等分成k份，用其中k-1份用来学习建立分类模型，用这个模型来测试剩下那一份，如此这般重复k次，使得每个例子测试且仅被测试一次，这种方法就叫做k重交叉验证。

#### 2.3.4 机器学习的应用

机器学习应用广泛，无论是在军事领域还是民用领域，都有机器学习算法施展的机会，主要包括以下几个方面。

（1）数据分析与挖掘

（2）模式识别

（3）在生物信息学上的应用

（4）更广阔的领域

## 3 数据集合的定义

Weka支持很多文件格式，包括arff、xrff、csv，甚至有libsvm的格式。其中arff是最常见的格式，arff是最常见的格式，我们在这里仅介绍这一种。arff全称是Attribute-Relation File Format，以下是一个arff格式的文件的例子。

文件中，“%”开头的是注释。剩余的可以分为两大部分，头信息（header information）的数据信息（data information）。

头信息中，“@relation”开头的行代表关系名称，在整个文件的第一行(除去注释)。格式是@relation;“@attribute”开头的代表特征，格式是@attribute；“@data”是实际的数据部分。每一行代表一个实例，可以认为是一个特征向量。各个特征的顺序与头信息中attribute逐个对应，特征值之间用逗号分割。在有监督分类中，最后一列是标注的结果。某些特征的数值如果是缺失的，可以用“?”代替。

Weka所处理的数据集是一个二维的表格。在3.1-3.4中的代码距举例，我们将使用Weka自带的weather数据进行讲解。

表3-1 weather数据集合

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 序号 | outlook | temperature | humidity | windy | play |
| 1 | sunny | hot | high | FALSE | no |
| 2 | sunny | hot | high | TRUE | no |
| 3 | overcast | hot | high | FALSE | yes |
| 4 | rainy | mild | high | FALSE | yes |
| 5 | rainy | cool | normal | FALSE | yes |
| 6 | rainy | cool | normal | True | no |
| 7 | overcast | cool | normal | True | yes |
| 8 | sunny | mild | high | FALSE | no |
| 9 | sunny | cool | normal | FALSE | yes |
| 10 | rainy | mild | normal | FALSE | yes |
| 11 | sunny | mild | normal | TRUE | yes |
| 12 | overcast | mild | high | TRUE | yes |
| 13 | overcast | hot | normal | FALSE | yes |
| 14 | rainy | mild | high | TRUE | no |

### 3.1 术语

表格里的一个横行称作一个实例（Instance），相当于统计学中的一个样本，或者数据库中的一条记录。竖行称作一个属性（Attribute），相当于统计学中的一个变量，或者数据库中的一个字段。这样一个表格，或者叫数据集，在Weka看来，呈现了属性之间的一种关系（Relation）。

### 3.2 声明

声明分为关系声明和属性声明。

#### 3.2.1 关系声明

关系名称在arff文件的第一个有效行来定义，格式为

@relation <relation-name>

<relation-name>是一个字符串。如果这个字符串包含空格，它必须加上引号（指英文标点的单引号或双引号）。

#### 3.2.2 属性声明

属性声明用一列以“@attribute”开头的语句表示。数据集中的每一个属性都有它对应的“@attribute”语句，来定义它的属性名称和数据类型。

这些声明语句的顺序很重要。首先它表明了该项属性在数据部分的位置。例如，“humidity”是第三个被声明的属性，这说明数据部分那些被逗号分开的列中，第三[列数据](https://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=74633135&ss_c=ssc.citiao.link)85 90 86 96 ... 是相应的“humidity”值。其次，最后一个声明的属性被称作class属性，在分类或回归任务中，它是默认的目标变量。

属性声明的格式为：

@attribute <attribute-name> <datatype>

其中<attribute-name>是必须以字母开头的字符串。和关系名称一样，如果这个字符串包含空格，它必须加上引号。

Weka支持的<datatype>有四种，分别是：

numeric-------------------------数值型

<nominal-specification>-----分类（nominal）型

string----------------------------字符串型

date [<date-format>]--------日期和时间型

其中<nominal-specification>和<date-format>将在下面说明。还可以使用两个类型“integer”和“real”，但是Weka把它们都当作“numeric”看待。注意 “integer”，“real”，“numeric”，“date”，“string”这些关键字是区分大小写的，而“relation”, “attribute”和“data”则不区分。

### 3.3 属性

属性大致分为五类

#### 3.3.1 数值属性

数值型属性可以是整数或者实数，但Weka把它们都当作实数看待。

#### 3.3.2 分类属性

分类属性由<nominal-specification>列出一系列可能的类别名称并放在花括号中：{<nominal-name1>,<nominal-name2>,<nominal-name3>,...}。数据集中该属性的值只能是其中一种类别。

例如如下的属性声明说明“outlook”属性有三种类别：“sunny”，“ overcast”和“rainy”。而数据集中每个实例对应的“outlook”值必是这三者之一。

@attribute outlook {sunny, overcast, rainy}

如果类别名称带有空格，仍需要将之放入引号中。

#### 3.3.3 字符串属性

字符串属性中可以包含任意的文本。这种类型的属性在文本挖掘中非常有用。

示例：

@ATTRIBUTE LCC string

#### 3.3.4 日期和时间属性

日期和时间属性统一用“date”类型表示，它的格式是

@attribute <name> date [<date-format>]

其中<name>是这个属性的名称，<date-format>是一个字符串，来规定该怎样解析和显示日期或时间的格式，默认的字符串是ISO-8601所给的日期时间组合格式“yyyy-MM-ddTHH:mm:ss”。

数据信息部分表达日期的字符串必须符合声明中规定的格式要求。

#### 3.3.5 Relational型属性

在Weka3.5版中增加了一种属性类型叫做Relational，有了这种类型我们可以像[关系型数据库](https://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=277136&ss_c=ssc.citiao.link)那样处理多个维度了。但是这种类型目前还不见广泛应用，暂不作介绍。

### 3.4 数据

一般的数据分为数据信息和稀疏数据

#### 3.4.1 数据信息

数据信息中“@data”标记独占一行，剩下的是各个实例的数据。

每个实例占一行。实例的各属性值用逗号“,”隔开。如果某个属性的值是缺失值（missing value），用问号“?”表示，且这个问号不能省略。例如：

@data

sunny,85,85,FALSE,no

?,78,90,?,yes

字符串属性和分类属性的值是区分大小写的。若值中含有空格，必须被引号括起来。例如：

@relation LCCvsLCSH

@attribute LCC string

@attribute LCSH string

@data

AG5, 'Encyclopedias and dictionaries.;Twentieth century.'

AS262, 'Science -- Soviet Union -- History.'

日期属性的值必须与属性声明中给定的相一致。例如：

@RELATION Timestamps

@ATTRIBUTE timestamp DATE "yyyy-MM-dd HH:mm:ss"

@DATA

"2001-04-03 12:12:12"

"2001-05-03 12:59:55"

#### 3.4.2 稀疏数据

有的时候数据集中含有大量的0值（比如购物篮分析），这个时候用稀疏格式的数据存贮更加省空间。

稀疏格式是针对数据信息中某个实例的表示而言，不需要修改arff文件的其它部分。看如下的数据：

@data

0, X, 0, Y, "class A"

0, 0, W, 0, "class B"

用稀疏格式表达的话就是

@data

{1 X, 3 Y, 4 "class A"}

{2 W, 4 "class B"}

每个实例用花括号括起来。实例中每一个非0的属性值用<index> <空格> <value>表示。<index>是属性的序号，从0开始计；<value>是属性值。属性值之间仍用逗号隔开。这里每个实例的数值必须按属性的顺序来写，如{1 X, 3 Y, 4 "class A"}，不能写成{3 Y, 1 X, 4 "class A"}。

注意在稀疏格式中没有注明的属性值不是缺失值，而是0值。若要表示缺失值必须显式的用问号表示出来。

### 3.5 csv转换为arff格式

我们通常接触到的数据为csv格式较多，若想要将数据更改为Weka通用的arff格式，我们可以利用Weka工具进行格式的转换。

#### 3.5.1 csv格式

逗号分隔值（Comma-Separated Values，CSV，有时也称为字符分隔值，因为分隔字符也可以不是逗号），其文件以纯文本形式存储表格数据（数字和文本）。纯文本意味着该文件是一个字符序列，不含必须像二进制数字那样被解读的数据。CSV文件由任意数目的记录组成，记录间以某种换行符分隔；每条记录由字段组成，字段间的分隔符是其它字符或字符串，最常见的是逗号或制表符。通常，所有记录都有完全相同的字段序列。通常都是纯文本文件。建议使用Word或是记事本来开启，再则先另存新档后用Excel开启，也是方法之一[4]。

#### 3.5.2 arff格式

arff是一种Weka专用的文件格式，由Andrew Donkin创立，有传言说arff代表Andrew's Ridiculous File Format(安德鲁的荒唐文件格式)，但在Weka的正式文档中明确说明arff代表Attribute-Relation File Format(属性——关系文件格式)。该文件是ASCII文本文件，描述共享一组属性结构的实例列表，由独立且无序的实例组成，是Weka表示数据集的标准方法，arff不涉及实例之间的关系。

#### 3.5.3 获取csv文件

本次案例我们使用UCI网站中的Soybean(Large)数据进行演示，我们先从网站中下载下来我们所需要的数据，进入网站后选择Download按钮进入下载页面，我们选择soybean-large.data数据进行下载下来发现是一个纯文本文件，用记事本打开，发现此文件大约有300多条实例，属性有36个我们只在这里介绍一下各个属性的含义。

将网站主页中的属性名（features）依次复制到该data文件中，属性之间用逗号隔开，这样我们就构造出来了csv格式的数据文件。保存文件，将该文件的后缀改为csv，我们就得到这个数据集的csv格式文件。

（1）class：种类，为字符串型，表示改大豆被分为的种类；

（2）date：生长时间，数字型，范围在0-10；

（3）plant-stand：植物架子的高度，数字型，范围在0-1；

（4）precip：降雨量，数字型，范围在0-1；

（5）temp：温度，数字型，范围在0-1；

（6）hail：冰雹，数字型，范围在0-1；

（7）crop-hist：裁剪历史记录，数字型，范围在0-1；

（8）area-damaged：受损区域，数字型，范围在0-1；

（9）severity：严重程度，数字型，范围在0-1；

（10）seed-tmt：种子健康程度，数字型，范围在0-1；

（11）germination：发芽情况，数字型，范围在0-1；

（12）plant-growth：植物生长状况，数字型，范围在0-1；

（13）leaves：叶子情况，数字型，范围在0-1；

（14）leafspots-halo：叶斑晕，数字型，范围在0-1；

（15）leafspots-marg：叶斑边缘部分，数字型，范围在0-1；

（16）leafspot-size：叶斑大小，数字型，范围在0-1；

（17）leaf-shread：叶鼩，数字型，范围在0-1；

（18）leaf-malf：叶子大小，数字型，范围在0-1；

（19）leaf-mild：叶温和，数字型，范围在0-1；

（20）stem：茎，数字型，范围在0-1；

（21）lodging：所在环境，数字型，范围在0-1；

（22）stem-cankers：茎溃疡情况，数字型，范围在0-1；

（23）canker-lesion：溃疡病损，数字型，范围在0-1；

（24）fruiting-bodies：子实体，数字型，范围在0-1；

（25）external decay：外部腐朽，数字型，范围在0-1；

（26）mycelium：菌丝体，数字型，范围在0-1；

（27）int-discolor：内部脱色，数字型，范围在0-1；

（28）sclerotia：菌核，数字型，范围在0-1；

（29）fruit-pods：果荚，数字型，范围在0-1；

（30）fruit spots：水果斑点，数字型，范围在0-1；

（31）seed：种子，数字型，范围在0-1；

（32）mold-growth：霉菌生长，数字型，范围在0-1；

（33）seed-discolor：种子变色，数字型，范围在0-1；

（34）seed-size：种子大小，数字型，范围在0-1；

（35）shriveling：皱缩，数字型，范围在0-1；

（36）roots：根，数字型，范围在0-1。

#### 3.5.4 转换csv格式为arff格式

步骤大致分为三步：

（1）打开Weka，选择explorer模式

（2）这时会打开新的explorer窗口，我们选择Open File选项，可以对我们的数据进行选择。

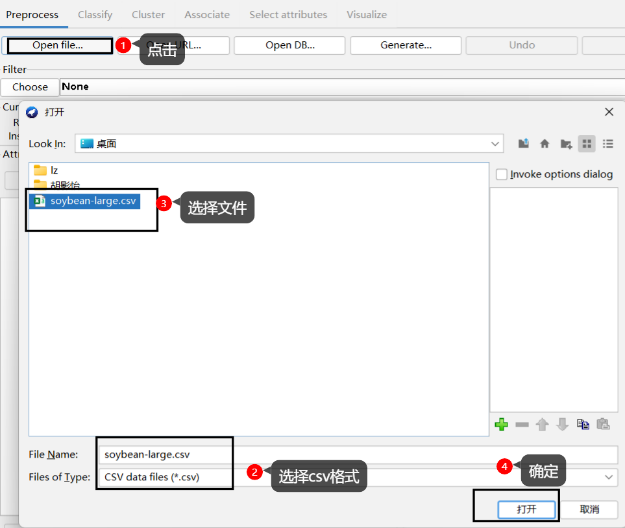


图3.1 在Weka中打开本地数据

（3）打开文件后，我们点击save选择文件类型为arff，选择文件名称以及存放位置后，点击保存，这样我们的数据就保存为了arff格式



图3.2 在Weka中另存为arff格式数据

## 4 研究的目的

如今是如文时代早已悄然过去，视频时代也已经到来，不论是电视剧，还是短视频，各种类型的视频资源已然称为当代年轻人的消遣时光的资源。本次探索就是仅聚焦B站视频平台，对其科技区2019年1月到2020年3月的公开数据进行爬取，并进行深度解析，分析一些优质的UP主，进行推荐[5]。

### 4.1 视频时代到来的原因

互联网大潮之下，内容行业发生了巨大的变化，主流内容渠道、传播介质出现置换，新媒体行业发展愈加成熟，并开始从图文阶段过渡至视频时代。视频于用户端的渗透在广度、深度方面均有明显体现。全民视频时代已至，超九成网民是视频用户。

#### 4.1.1 对视频用户的分析

最近几年随着视频平台的，视频化兴趣社区成为年轻人主要的内容消费场域，截至2020年6月，中国网民规模为9.4亿，而网络视频用户规模达8.9亿，即94.5%的网民是网络视频用户。而在年轻用户中，他们对视频的粘性也明显增长。据Quest Mobile平台数据显示，2020年9月，移动视频是90后~00后用户月人均使用时长位居第一、第二的应用类型，且同比增长率明显高于社交和游戏。

#### 4.1.2 5G技术的发展

5G时代到来，技术壁垒的突破为视频媒介传播提供了良好的发展空间。技术加持下，视频信息含量、视听体验等原生优势得以完整呈现。

视频信息含量丰富，信息展现具体直观﹐动态的视频能够承载着更多信息内容﹐也可以提供更为具象、沉浸、生动的感官体验﹐为互动、娱乐带来新可能。

视频能够提供良好的视觉听觉体验。一方面，用户多感官接收信息﹐能够有效降低理解门槛;另一方面﹐随着4K/8K、VR/AR等技术的运用﹐视觉听觉体验不断升级﹐视频将在感官精神方面带来更佳的用户体验。

信息技术发展和原生优势的共同作用下﹐视频成为满足用户多样场景需求的重要工具。

#### 4.1.3 社交因素

兴趣取代传统社交要素，成为新一代年轻人建立认同感和信任感的纽带，视频则是他们最偏爱的兴趣内容载体。

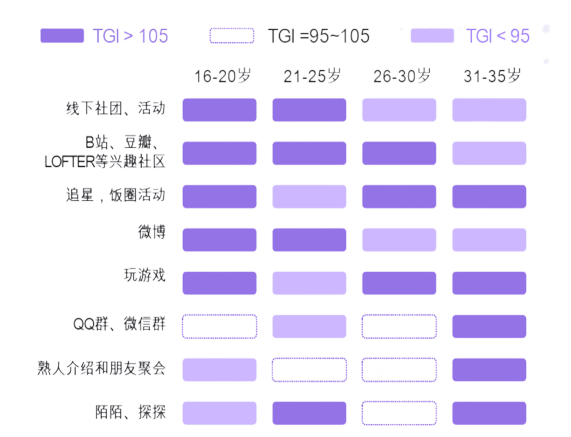


图4.1 各年龄群体更偏好认识新朋友的途径

网生一代年轻人更热衷于通过兴趣内容判断彼此的审美、价值观的吻合度，此为建立认同感的基础,结交新的朋友。因此，一代年轻人更偏爱通过兴趣类的社群、社区和平台认识新朋友。

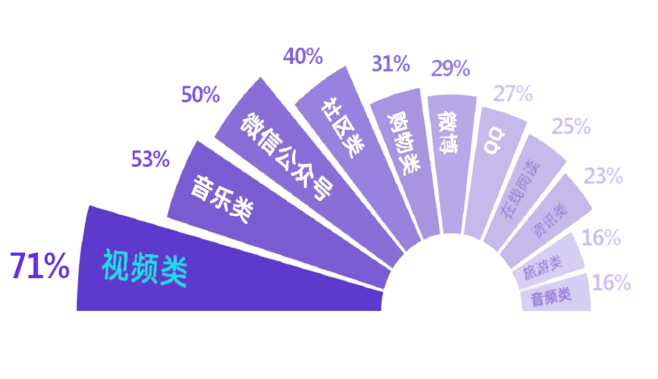


图4.2 年轻人培养和满足自己的兴趣爱好的平台

### 4.2 创作门槛的降低

新一代年轻人普遍利用视频创作释放表达欲以兴趣传递自我表达

随着网络的迅速发展、早期的流量红利以及移动智能手机的普及，视频剪辑编辑应用进入市场，低了视频生产的门槛。创作者通过手机拍摄后﹐可直接进行剪辑，大量UGC（UGC是一种互联网的术语，是指广大用户将自己生成的原创内容分享给朋友和他人，它是用户使用互联网的一种新的方式，即下载和上传并重的分享方式）内容得到生产和社交传播。

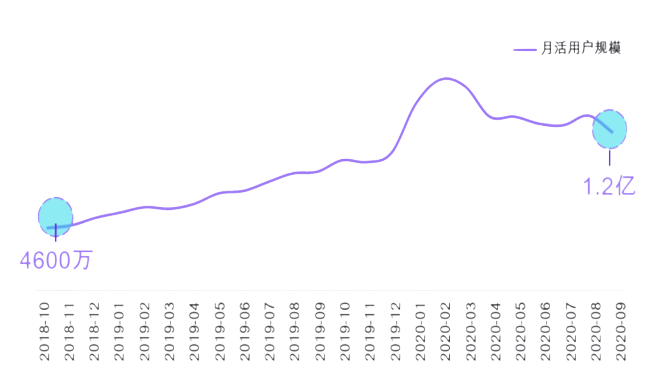


图4.3 视频工具月活用户增长情况

内容爆发、红利见顶，用户对UGC的内容专业性提出更高要求，随着视频行业成熟化发展，用户对内容专业性的要求提升明显﹐促使专业用户生产的PUGC（全称：Professional Generated Content + User Generated Content，即“专业用户生产内容”或“专家生产内容”）内容得到更多关注。

兴趣是年轻人的沟通语言,能够激发更强的情感共鸣和认同感、以此作为建立社交关系的基础。与此同时，年轻人对高品质的UGC内容需求日益增长，而聚集高品质创作者内容的视频社区，则成为年轻人主要的内容消费聚集场域和自我表达的阵地。

### 4.3 选取B站的原因

移动互联网的发展和视频制作的便捷化让每一个人都有机会成为创作者，表达自我、展现创意﹐他们以视频为载体，解构与重构流行文化。

视频社区B站聚集了大量优质的年轻一代内容创作者，他们被用户称为“UP主”，UP主们积极创作，产出优质PUGV内容达B站整体播放量的91%。

2020年Q3,B站月活跃UP主同比增长51%,达到170万，月均投稿量同比增长79%,达560万。以视频为核心用户在B站源源不断创作高质量的兴趣内容。

B站已发展成为年轻人高度聚集的文化社区和视频平台。B站拥有7000多个文化圈层。2019年10月至2020年9月，B站播放内容标签超过1000万个，同比上一年增加43%；同时，用户人均每月观看超过20个分区的兴趣内容，展现出年轻一代兴趣跨圈层趋势。年轻人呈现兴趣多元化、跨圈层趋势社区汇集成千上万的文化兴趣圈层。同时美食、时尚、知识、生活、数码等年轻人喜欢的内容在B站增长显著。

#### 4.3.1 美食分类

B站的美食类覆盖美食的全方位品类和场景，满足下厨、探店、陪伴等多种需求。

B站美食类内容覆盖面广泛，包括美食制作、美食探店、美食测评、美食记录等,观众可以学习厨艺或种草美食;而田园美食中则包括古风、田园等日常生活罕见场景﹐拓展了美食的边界。

#### 4.3.2 时尚分类

B站时尚分区细分多元且专业性强，覆盖各种场景和人群对美妆、穿搭和健身的内容需求。

B站时尚类分区内包含美妆、健身、服饰等二级分区，其中,“美妆”包含仿妆、国风美妆等多个个性化特色圈层；“健身”多是打卡挑战、干货内容;“服饰”中的汉服圈则是吸引了众多同好。

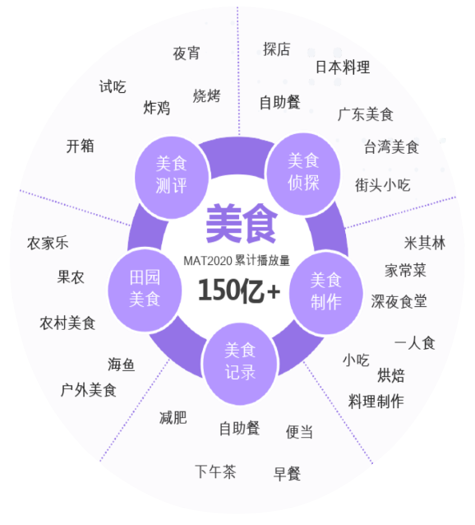


图4.4 美食类分类饼状图

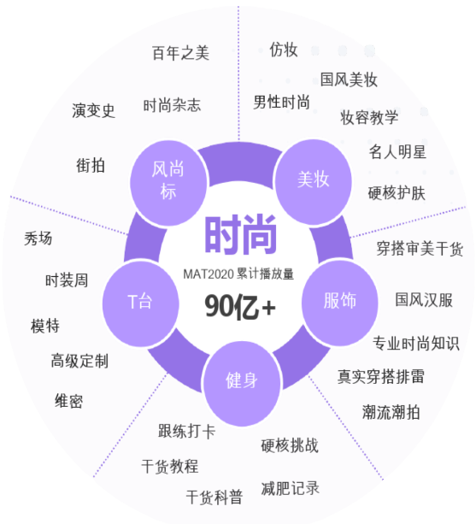


图4.5 时尚类分类饼状图

#### 4.3.3 知识分类

B站知识分类兼备专业性和趣味性，成为年轻人偏爱的知识获取渠道，这其实都得益于UP主产出的高品质且有趣味的内容、良好的社区交流氛围、年轻用户的强烈学习需求，B站泛知识类内容蓬勃发展。

2019年4月17日，央视网文章《知道吗﹖这届年轻人爱上B站搞学习》将B站的学习生态带入大众视野﹐引发网络热议。随后﹐罗翔说刑法、毕导THU等UP主全网走红，使得B站成为新一代“学习”网站。

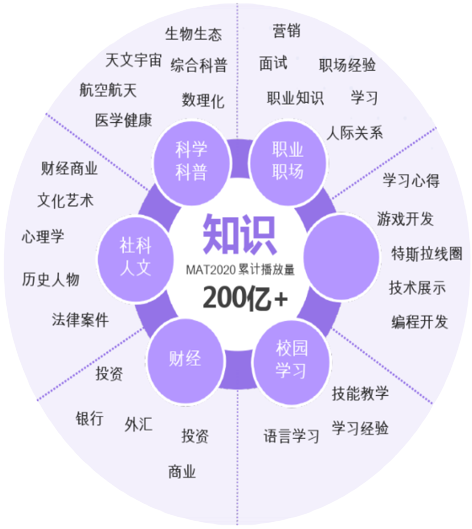


图4.6 知识类分类饼状图

#### 4.3.4 生活分类

Vlog已成为年轻人记录生活分享日常的常用形式。除了分享个人日常外，还有UP主以动物为主角展开创作，如UP主奶糕成精档案社，高质量猫片频受认可，多次登上全站排行榜。Vlog、萌宠、汽车等生活类视频受关注，生活区成年轻人记录日常的主要阵地之一。

#### 4.3.5 数码分类

B站数码区内容类型以数码科普、评测体验为主，数码科普、评测体验等优质UP主内容频频引爆出圈，内容硬核且有趣。2019年6月，B站知名UP“老师好我叫何同学”的5G日常体验为主题的视频引爆社交网络﹐收获大量粉丝﹐同时带动了大众对于数码科技内容领域的关注。

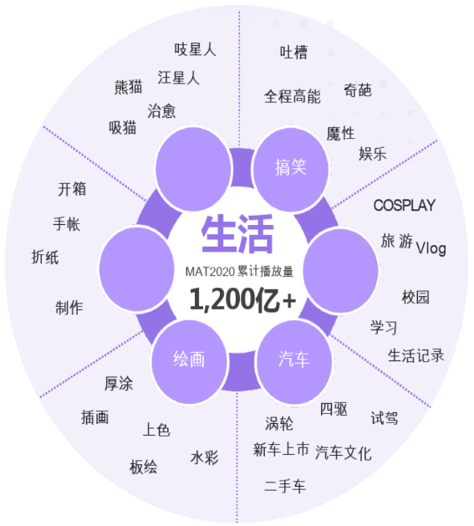


图4.7 生活类分类饼状图

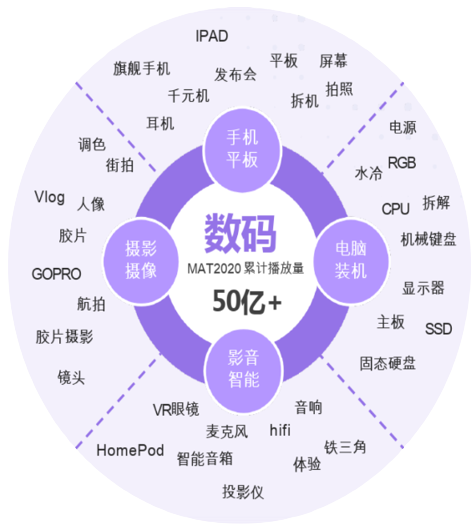


图4.8 数码类分类饼状图

#### 4.3.6 总结

视频行业已从早期野蛮生长的“草莽时代"发展到现在的良序善治的合规时代。与此同时，视频创作者群体将向专业化、垂直化、多元化、精细化靠拢。

（1）内容专业化

随着视频行业发展和用户需求升级，视频创作向专业化发展，主要体现在两方面：一是视频创作者业余兼职向专业全职发展；二是创作者生产内容向专业化发展，包括内容本身的专业化和内容制作流程的专业化。

（2）领域垂直化

行业成熟化发展之下，视频意义不再局限于娱乐大众，还包括获取知识、兴趣社交等方面。知识、美食等垂直领域创作者受到关注，小众特色创作者也有一定规模的粉丝群体。视频创作者生态形成泛娱乐垂直共繁荣局面。

（3）主体多元化

视频创作者类型向多元化发展，明星、机构等多元背景的群体成为视频创作者。

（4）运营精细化

视频创作者在内容制作、内容分发、粉丝交互等方面进行精细化运营，以实现差异化发展，从而增强行业竞争力。

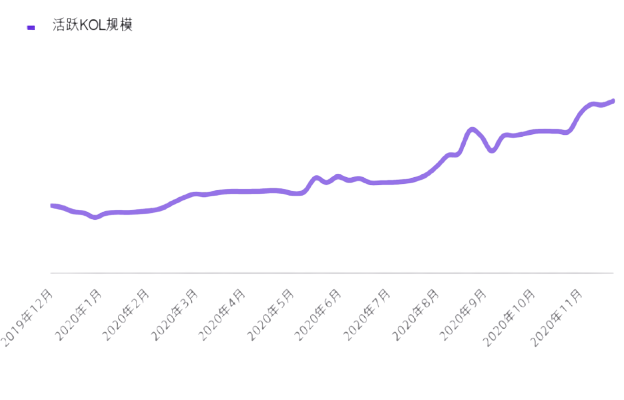


图4.9 主流平台活跃KOL规模趋势

## 5 对数据的清洗和处理

在进行实验前，需要了解数据的相关信息，下文则详细介绍了论文所需的数据集的网址来源，以及数据集内整体数据和单个数据的属性和特点。

B站播放量数据集合：选取2019年1月-2020年3月，B站播放量过五万的视频的一些信息，主要特征有分区、视频编号、作者名、作者id、视频的发布日期、视频链接、视频标题、投币数、总弹幕数、收藏数、点赞数、总评论数、分享数以及总播放量，共计50,130行数据。该数据集来源于和鲸社区，数据集的网址为：https://www.heywhale.com/mw/dataset/5f6dbf6071c70000307 f0e60。

### 5.1 Jupyter Notebook介绍

Jupyter Notebook是以网页的形式打开，可以在网页页面中直接编写代码和运行代码，代码的运行结果也会直接在代码块下显示。如在编程过程中需要编写说明文档，可在同一个页面中直接编写，便于作及时的说明和解释[6]。

#### 5.1.1 IPython介绍

IPython是一个python的交互式shell，比默认的python shell好用得多，支持变量自动补全，自动缩进，支持bash shell命令，内置了许多很有用的功能和函数。

IPython是基于BSD开源的。

IPython为交互式计算提供了一个丰富的架构，包含：

（1）强大的交互式shell

（2）Jupyter内核

（3）交互式的数据可视化工具

（4）灵活、可嵌入的解释器

（5）易于使用，高性能的并行计算工具

#### 5.1.2 组成部分

Jupyter Notebook由网页应用和文档两者共同

（1）网页应用：网页应用即基于网页形式的、结合了编写说明文档、数学公式、交互计算和其他富媒体形式的工具。简言之，网页应用是可以实现各种功能的工具。

（2）文档：即Jupyter Notebook中所有交互计算、编写说明文档、数学公式、图片以及其他富媒体形式的输入和输出，都是以文档的形式体现的。

这些文档是保存为后缀名为.ipynb的JSON格式文件，不仅便于版本控制，也方便与他人共享。

此外，文档还可以导出为：HTML、LaTeX、PDF等格式，这是我认为最方便的一点，便于文件内容的传输。

#### 5.1.3 pandas介绍

本次对数据的清洗工作全部使用pandas这个科学计算库进行。

Pandas是python的一个数据分析包，最初由AQR Capital Management于2008年4月开发，并于2009年底开源出来，目前由专注于Python数据包开发的PyData开发team继续开发和维护，属于PyData项目的一部分。Pandas最初被作为金融数据分析工具而开发出来，因此，pandas为时间序列分析提供了很好的支持。 Pandas的名称来自于面板数据（panel data）和python数据分析（data analysis）。panel data是经济学中关于多维数据集的一个术语，在Pandas中也提供了panel的数据类型。

数据结构

Series：一维数组，与Numpy中的一维array类似。二者与Python基本的数据结构List也很相近。Series如今能保存不同种数据类型，字符串、boolean值、数字等都能保存在Series中。

Time- Series：以时间为索引的Series。

DataFrame：二维的表格型数据结构。很多功能与R中的data.frame类似。可以将DataFrame理解为Series的容器。

Panel ：三维的数组，可以理解为DataFrame的容器。

Panel4D：是像Panel一样的4维数据容器。

PanelND：拥有factory集合，可以创建像Panel4D一样N维命名容器的模块。

### 5.2 数据概览

图5.1为视频信息表



图5.1 部分数据截图

对数据进行简单分析后会发现，这里面有很多空数据、重复以及无效数据，下面我将使用基于IPython的Jupyter Notebook对数据进行清洗，并重新保存，避免在进行数据分析的时候这些干扰项干扰我们的判断。

#### 5.2.1 各字段数量

import pandas as pd

data = pd.read\_csv("./科技区2019-1\_2020-3.csv")

data.shape

data.info()



图5.2 对数据的查看

#### 5.2.2 缺失值的数量：

data.isnull().sum()



图5.3 每个属性缺失值的数量

### 5.3 数据清洗

接下来对数据进行去重、去空值等操作

#### 5.3.1 数据去重

data.drop\_duplicates(inplace=True)

data.info()

可以看到此时数据已经减少了1312条数据

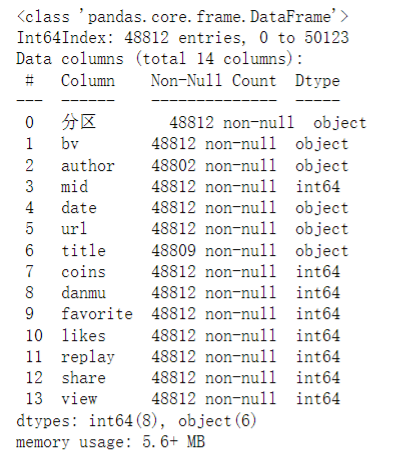


图5.4 去重后的数据

#### 5.3.2 删除数据缺失值

data.dropna(inplace=True)

data.info()

可以看到此时数据减少了13条

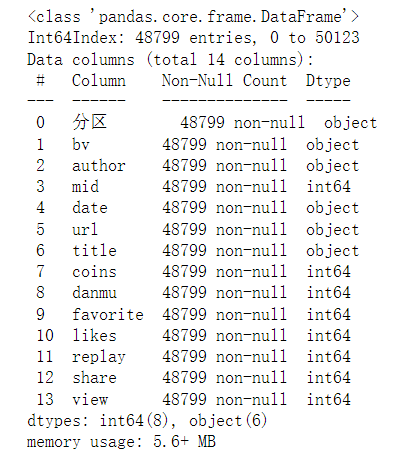


图5.5 删除缺失值后的数据

### 5.4 ILF模型的构建

RFM模型是衡量客户价值和创利能力的重要工具和手段。通过一个客户近期购买行为、购买的总体频率以及消费金额三项指标来描述客户的价值状况。

R：最近一次消费时间（最近一次消费到参考时间的间隔）

F：消费的频率(消费了多少次）

M：消费的金额 （总消费金额）

但RFM模型并不能评价视频的质量，所以在这里针对up主的视频信息构建了IFL模型，以评估视频的质量。

#### 5.4.1 对IFL模型的解释

I值：反映的是平均每个视频的互动率，互动率越高，表明其视频更能产生用户的共鸣，使其有话题感。

F值表示的是每个视频的平均发布周期，每个视频之间的发布周期越短，说明内容生产者创作视频的时间也就越短，创作时间太长，不是忠实粉丝的用户可能将其遗忘。

L值表示的是统计时间内发布视频的平均点赞率，越大表示视频质量越稳定，用户对up主的认可度也就越高。

计算方法：

接下来我们将利用pandas这一科学计算工具对I、F、L三值进行计算，并保存为csv格式数据，以便用于Weka的数据分析。

我们在对数据进行清洗后，还要对单个作者的作品总数进行统计，当数量太小，例如小于5的时候，就再次删除该作者，因为该作者及其作品不具有统计意义。

# 筛选视频数大于5

v\_count = data.groupby('author')['bv'].count()

# 将筛选出来的数据用data2进行保存

data2 = data[-data['author'].isin(v\_count[v\_count.values < 5].index)]

data2.index = range(len(data2))

此时的数据已经减少到了38568行，接下来我们对常用的计算三个值常用的变量保存起来。

# 总弹幕

danmu\_author = data2.groupby('author')['danmu'].sum()

# 总评论

replay\_author = data2.groupby('author')['replay'].sum()

# 总播放量

view\_author = data2.groupby('author')['view'].sum()

# 统计范围内的视频数量

vedio\_author = data2.groupby('author')['bv'].count()

#### 5.4.2 分别计算IFL的值

接下来我们继续利用pandas对I、F、L三个值进行计算

（1）计算I值：平均每个视频的互动率

# round对数据进行四舍五入

I = round((danmu\_author + replay\_author) / view\_author / vedio\_author \* 100, 2).reset\_index()

I.rename(columns={0:'I'}, inplace=True)

（2）计算F值：每个视频平均发布周期

# 将date 转变为时间格式

data2['date'] = pd.to\_datetime(data2['date'])

# 最早最晚发布时间

latest\_date = data2.groupby('author')['date'].max()

earliest\_date = data2.groupby('author')['date'].min()

F = (difference / vedio\_author).reset\_index()

F.rename(columns={0:'F'}, inplace=True)

（3）计算L值：发布视频的平均点赞数

data2['L'] =(data2['likes'] + data2['coins'] \* 2 + data2['favorite'] \* 3 + data2['share']) / data2['view'] \* 100

L = round((data2.groupby('author')['L'].sum()

/ vedio\_author).reset\_index(), 2)

L.rename(columns={0:'L'}, inplace=True)

这样就完成对IFL的计算，我们将三个值分别放在I、F、L三个变量中。

#### 5.4.3 将三个值进行合并和导出

完成对IFL的求取，我们现在对IFL三个值进行合并。

IFL = pd.merge(I, F, on='author', how='inner')

IFL = pd.merge(IFL, L, on='author', how='inner')



图5.6 将IFL放在同一个表中

将三个变量合并完成后，接下来我们进行对数据的下载。

IFL.rename(columns={'I':'I\(平均每个视频的互动率\)'}, inplace=True)

IFL.rename(columns={'F':'F\(每个视频的平均发布周期\)'}, inplace=True)

IFL.rename(columns={'L':'L\(统计时间内发布视频的平均点赞率\)'}, inplace=True)

# 保存为csv格式数据

df = pd.DataFrame(IFL)

df.to\_csv("./data.csv", index=False)

上面的代码就是对数据属性的重命名，还有导出工作。打开当前文件夹后，我们可以发现data.csv文件就保存在了这里。

现在我们可以对该文件进行数据分析了。

## 6 对ILF模型进行数据分析

再将csv数据导入Weka的时候，可能会出现乱码的情况，这时候我们就需要进行解决这些乱码问题。

### 6.1 解决乱码问题

我们要解决乱码问题，首先要知道是哪里导致了乱码，我在这里是对两个地方进行了更改。

#### 6.1.1 改变csv文件的编码

首先，我将我的data.csv文件进行重新编码并保存，数据导出后默认为GBK，我将其改变为utf-8，这样提高了文件的通用性。

更改编码的方式有很多种，我这里使用的是VS Code进行对文件的重新编码，这里我们使用该软件打开data.csv文件，在软件下方找到我们编码方式的按钮，鼠标左键单击后，会出现两个选项，我们选择save with Encoding后，继续选择编码格式，找到utf-8后单击回车即可。

#### 6.1.2 改变Weka的默认编码打开方式

当我改变文件编码后，用Weka打开文件发现还是乱码，我百度后发现Weka软件的默认编码字符集为Cp1252。于是有对软件默认编码进行搜索，终于在一篇博客中找到了答案。

首先我们要找到Weka的安装目录，里面有一个RunWeka.ini的文件，用记事本打开。之后找到33行，将注释删除，添加这样的代码fileEncoding=utf-8。再次使用Weka打开csv文件后，发现问题已经解决，所有汉字全部可以正常显示了。

我们可以使用Weka自带的转换工具，将csv格式数据另存为arff格式数据，转换后便于我们的下次打开。详细操作步骤请看第3.5小结。

### 6.2 K-Means原理

聚类是一个将数据集中在某些方面相似的数据成员进行分类组织的过程，聚类就是一种发现这种内在结构的技术，聚类技术经常被称为无监督学习。

K均值聚类是最著名的划分聚类算法，由于简洁和效率使得他成为所有聚类算法中最广泛使用的。给定一个数据点集合和需要的聚类数目K，K由用户指定，K均值算法根据某个距离函数反复把数据分入K个聚类中。

K均值聚类算法，是一种迭代求解的聚类分析算法，其步骤是，预将数据分为K组，则随机选取K个对象作为初始的聚类中心，然后计算每个对象与各个种子聚类中心之间的距离，把每个对象分配给距离它最近的聚类中心。聚类中心以及分配给它们的对象就代表一个聚类。每分配一个样本，聚类的聚类中心会根据聚类中现有的对象被重新计算。这个过程将不断重复直到满足某个终止条件。终止条件可以是没有（或最小数目）对象被重新分配给不同的聚类，没有（或最小数目）聚类中心再发生变化，误差平方和局部最小[7]。

### 6.3 使用Weka进行分析

接下来我们将使用Weka中K-Means模型来对数据进行训练。

#### 6.3.1 对K值的选定

我们使用手肘法来进行对K的选定，手法的核心指标是SSE(sum of the squared errors，误差平方和)，手肘法的核心思想是:随着聚类数k的增大，样本划分会更加精细，每个簇的聚合程度会逐渐提高，那么误差平方和SSE自然会逐渐变小。并且，当k小于真实聚类数时，由于k的增大会大幅增加每个簇的聚合程度，故SSE的下降幅度会很大，而当k到达真实聚类数时，再增加k所得到的聚合程度回报会迅速变小，所以SSE的下降幅度会骤减，然后随着k值的继续增大而趋于平缓，也就是说SSE和k的关系图是一个手肘的形状，而这个肘部对应的k值就是数据的真实聚类数。当然，这也是该方法被称为手肘法的原因。

接下来我们使用Weka来对SSE进行判断，表6-1为结果。

表6-1 K值和SSE的选择

|  |  |
| --- | --- |
| K值 | SSE |
| 2 | 1661.6798358503997 |
| 3 | 1594.2107177362273 |
| 4 | 1580.8892633739301 |
| 5 | 1573.833483510359 |
| 6 | 1570.4724264952451 |

我们可以明显的看到当K为3的时候，SSE的下降速度是最快的，此时我们选择K为3进行模型的建立。

#### 6.3.2 对数据的预测

将K值设为3后，我们使用将全部数据分为两部分的方式对模型进行训练，将训练数据设置为60%，预测数据设置为40%，等待片刻后可以得到结果。我们将结果整理到表6-2中。

表6-2 K-Means进行预测的结果

|  |  |
| --- | --- |
| 编号 | 占比情况 |
| 0 | 8% |
| 1 | 58% |
| 2 | 34% |

### 6.4 结论

K-Means预测的结果和我们ILF模型构建结果之间的相互分析可以得到如下结论：

（1）索引为0的up主

该类别的I值略高于平均值，证明与观众的互动率较高，话题性也较高；但F值明显高于其他两类，创作周期较长；

L值平均点赞率良好，仍有可提升空间；该类别大部分为业余up主，处于成长期。

（2）索引为1的up主

该部分占了58%，是该分区的主要群体；该类别的I值稍低于平均值，与观众的互动良好；F值为三者中最低，更新频率高，活跃程度高；

但L值平均点赞率是短板，较低，暂未能获得大量观众的点赞、投币、收藏；需要提高视频稿件的质量，创作符合观众口味的视频，收获更多的三连

（3）索引为2的up主

该部分up主群体I值远高于平均值，受到观众的喜爱，话题感十足；视频更新稳定，在保证不拖更的同时，收获到观众大量的三连（点赞、投币、收藏）；

该类up主粉丝数目也大部分超过十万，能够提供高质量且稳定的视频稿件，为该分区的高价值up主群体。

## 参考文献

[1] 百度百科, URL: https://baike.baidu.com/item/weka [2022.6.16].

[2] Weka 官网, URL: https://www.cs.waikato.ac.nz/ml/weka [2022-6-16].

[3] 博客, URL: http://t.zoukankan.com/chamie-p-4551514.html [2022-6-16].

[4] 博客, URL: https://baike.baidu.com/item/CSV/10739 [2022-6-18].

[5] 豆丁网, URL: https://www.docin.com/p-2608559360.html [2022-6-20]

[6] 百度百科, URL https://baike.baidu.com/item/Jupyter/20423051 [2022-6-21]

[7] 百度百科, URL https://baike.baidu.com/item/K均值聚类算法/15779627 [2022-6-22]