指标的数据收集方案

1. 指标的初步选取

下面是选取的12个影响票房因素的指标，根据数据的收集难度和量化的难度分成三个等级：

* 蓝色代表数据已收集，量化相对简单。
* 红色代表数据可收集，量化需求复杂。
* 绿色表示不知数据源从何获取，量化方案未知。

1. 导演的影响力（movie\_director）
2. 演员的影响力（movie\_actor）
3. 编剧的影响力（screenwriter）
4. 制作公司（production\_company）
5. 发行公司（distribution\_company）
6. 影片类型（movie\_type）
7. 档期（发行日期）（issue\_date）
8. 知识产权（intellectual\_property）
9. 制作技术（production\_technology）
10. 排片数据（row\_of\_tablets）
11. 网络口碑（reputation）
12. 点映（on\_demand）
13. 指标的具体量化方案

导演、演员、编剧、制作公司、发行公司、影片类型（15种类型单独量化）、知识产权、制作技术（3种单独量化）、排片、网络口碑量化后的指标为百分制。点映设置为虚拟变量0|1。下面介绍具体的数据收集和量化方法。

1. **导演的影响力**

储存导演的数据和信息需要单独去做一个数据表，因为导演不会因为已有的顺序而不发生变化。所以我们从豆瓣和艺恩电影库收集导演的信息、需要收集所需的数据表格。最能体现一个导演的水平就是导演的历史作品，初步决定收集导演的信息如下表：

表2-9 导演的信息收集表（18个指标细化信息）

| 指标名称 | 指标细化 | 数据采集源 | 备注说明 |
| --- | --- | --- | --- |
| 导演影响力 | 作品名称列表（冗余） | 中国票房  豆瓣电影  猫眼票房  时光网 | 导演实力影响因素 |
| 单部平均票房 |
| 票房排名 |
| 票房占比 |
| 历史最高票房   | 导演获奖数目 | | --- | | 导演提名数目 | | 擅长执导类型 | |
| 微博关注量 | 微博指数 | 导演的人气影响 |
| 微博转发量 |
| 微博评论量 |
| 微博整体指数 |
| 贴吧的关注量 | 百度指数  百度贴吧 |
| 贴吧的帖子数量 |
| 百度整体指数 |
| 豆瓣评分 | 豆瓣 |
| 豆瓣收藏人数 |
| 观众喜爱程度 | 时光网 |

1. **演员的影响力**

评价体系和导演的类似。

表2-10 演员信息收集

| 指标名称 | 指标细化 | 数据采集源 | 备注说明 |
| --- | --- | --- | --- |
| 演员影响力 | 代表作品名称 | 中国票房  豆瓣电影  猫眼票房  时光网 | 演员实力影响因素  权重为40% |
| 单部平均票房 |
| 票房排名 |
| 票房占比 |
| 历史最高票房   | 获奖数目 | | --- | | 提名数目 | | 擅长表演的类型 | |
| 微博关注量 | 微博指数 | 演员的人气影响  权重为60% |
| 微博转发量 |
| 微博评论量 |
| 微博整体指数 |
| 贴吧的关注量 | 百度指数  百度贴吧 |
| 贴吧的帖子数量 |
| 百度整体指数 |
| 豆瓣评分 | 豆瓣 |
| 豆瓣收藏人数 |
| 观众喜爱程度 | 时光网 |

1. **编剧的影响力**

表2-11 编剧的详细量化

| 指标名称 | 指标细化 | 备注 |
| --- | --- | --- |
| 编剧 | 编剧代表作品 | 豆瓣 |
| 微博关注量 | 微博和微博指数 |
| 微博转发量 |
| 微博评论量 |
| 微博整体指数 |
| 贴吧关注量 | 百度和百度指数 |
| 贴吧帖子数量 |
| 百度整体指数 |
| 豆瓣评分 | 豆瓣 |
| 观众喜爱程度 | 时光网 |

**4. 发行公司和制作公司：**根据2-8市场法则，20%发行公司可能占有市场份额的80%，在我国前十的大发行公司占据了80%的市场票房。所以从最开始我们将发行公司进行分类：大发行公司、中档发行公司和普通发行公司（因为不可能收集到所有的发行公司的信息，所以会将没有记录的发行公司归为普通发行公司）。根据对电影市场的调查和已收集到的数据对电影发行公司进行分类和量化。

表2-3 发行公司类别

|  |  |  |
| --- | --- | --- |
| 发行公司类别 | 发行公司名称 | 初步估量值 |
| 大发行公司 | 中影发行 | 市场占比约50% |
| 华夏发行 |
| 普通发行公司 | 光线影业 | 市场占比约30% |
| 博纳影业 |
| 万达影视 |
| 乐视影业 |
| 华谊兄弟 |
| 嘉映影业 |
| 其他发行公司 | - | 市场占比约20% |

表2-4 发行公司详细量化

| 指标名称 | 指标细化 | 备注 |
| --- | --- | --- |
| 发行公司 | 历史作品平均票房 | 可计算 |
| 发行公司类别 | 参照表2-3 |

使用计算公式： 量化后的值 = 发行过主要的电影票房的平均票房\*权重+发行公司类别\*权重。制作公司的量化方式与发行公司类似，因为很多的发行公司也是制作公司，公司列表可能不同，使用统一计算公式并且加权求得量化后的值，即计算票房占比。

**5. 影片类型：**根据获取到的数据，将在此将影片类型总共分成15类：类型可以进行单独的量化分析（设置虚拟变量0|1，表示某类型的有无），由于电影大多数都不是一种类型，所以先抽取主要类型进行分析。（备注：怎么抽取复杂类型，进行量化分析。）

表2-4 电影类型

| 指标名称 | 细化分类 | 备注 | 标识符 |
| --- | --- | --- | --- |
| 影片类型 | 喜剧 | 分类参考电影网站 | 依次整型1-15的值 |
| 悲剧 |
| 爱情 |
| 动画 |
| 文艺 |
| 青春 |
| 战争 |
| 悬疑 |
| 惊悚 |
| 动作 |
| 枪战 |
| 科幻 |
| 家庭 |
| 犯罪 |
| 魔幻 |

**6. 发行日期：**发行日期也就是档期，可以根据日期转换成档期。档期的量化公式，

档期的增益比 = （所在档期的时间段内的单部电影的平均票房 / 年度单部电影平均票房）

特定档期的权重 = 特定增益比 / 总的增益比

表2-5 档期分类表

| 指标名称 | 分类 | 备注 | 标识符 |
| --- | --- | --- | --- |
| 档期 | 贺岁档 | 收集电影的发行日期，根据发行日期进行档期分类。 | 5   | 4 | | --- | | 3 | | 2 | | 1 |   0 |
| 暑期档 |
| 国庆档 |
| 五一档 |
| 其余节假日档 |
| 普通档 |

**7. 制作技术：**比较同期上映的电影使用不同技术对票房的影响

表2-6 制作技术分类

| 指标名称 | 分类 | 备注 | 标识符 |
| --- | --- | --- | --- |
| 制作技术 | IMAX 3D | 已收集 | | 2 | | --- | | 1 |   0 |
| 3D |
| 2D |

**8. 排片数据：**主要从猫眼票房网站动态获取以下数据：未来10-15天内的场次占比数据、场次数目、上座率、场均人数。从猫眼的数据可以看出热门电影才有排片数据，大部分电影的排片数据是不确定的，所以在建立数据库的时候允许数据缺失（其他的排片）。

表2-7 排片数据

| 指标名称 | 指标细化 | 数据采集源 |
| --- | --- | --- |
| 排片数据 | 场次数目 | 猫眼票房 |
| 拍片占比 |
| 上座率 |
| 上映天数 |

**9. 点映：**有则为1，无则为0，可为空值。（真想把这个指标删除，因为历史数据很难获取，如果从当前开始可能会影响准确度，而且需要时间，如果有数据源就好。）老师写的文档描述如下：有市场号召力的影片做点映，口碑会呈几何倍数增长。对于艺术片也不失为好的宣传方法，通过少量银幕实现强口碑宣传，影响院线并扩大发行规模。

**10．网络口碑影响力：**主要以IMBD、豆瓣评分和**搜索指数：**使用百度指数和微博指数综合评分，动态更新，每天更新一次。综合指数和微博指数数据，百度指数和微博指数各占50%权重。

表2-12 网络口碑详细量化表

| 指标名称 | 指标细化 | 备注 | 数据采集源 |
| --- | --- | --- | --- |
| 网络口碑 | IMBD评分 | 其余评分不准确 | 豆瓣 |
| 豆瓣评分 |
| 近7天百度指数 | 百度指数  微博指数  取均值 | 百度指数 |
| 近30天百度指数 |  |
| 近7天增长幅度 |  |
| 三个月微博指数均值 | 微博指数 |
| 当月微博指数均值 |  |
| 当月增长幅度 |  |
| 想看指数 | 综合权威售票网站 |  |
| 同档期搜索率占比 | Google白皮书指标 | 计算可得 |
| 预告片点击量 | 权威视频网站 | 爱奇艺、优酷、PPTV、搜狐视频 |

**11. 知识产权影响力（IP）**

主要是怎么获取知识产权的数据？知识产权涉及的类型太多了，收集数据都有困难。

1. 缺失值处理相关知识（摘自网络）

在对缺失数据进行处理前，了解数据缺失的机制和形式是十分必要的。将数据集中不含缺失值的变量称为**完全变量**，数据集中含有缺失值的变量称为**不完全变量**。从缺失的分布来将缺失可以分为完全随机缺失，随机缺失和完全非随机缺失。[1]处理不完整数据集的方法主要有三大类：删除元组、数据补齐、不处理。

1. **删除元组**，也就是将存在遗漏信息属性值的对象（元组，记录）删除，从而得到一个完备的信息表。这种方法简单易行，在对象有多个属性缺失值、被删除的含缺失值的对象与初始数据集的数据量相比非常小的情况下非常有效，类标号缺失时通常使用该方法。 然而，这种方法却有很大的局限性。它以减少历史数据来换取信息的完备，会丢弃大量隐藏在这些对象中的信息。在初始数据集包含的对象很少的情况下，删除少量对象足以严重影响信息的客观性和结果的正确性；因此，当缺失数据所占比例较大，特别当遗漏数据非随机分布时，这种方法可能导致数据发生偏离，从而引出错误的结论。说明:删除元组，或者直接删除该列特征，有时候会导致性能下降。
2. **数据补齐，**这类方法是用一定的值去填充空值，从而使信息表完备化。通常基于统计学原理，根据初始数据集中其余对象取值的分布情况来对一个缺失值进行填充。数据挖掘中常用的有以下几种补齐方法：
3. **人工填写**（filling manually）由于最了解数据的还是用户自己，因此这个方法产生数据偏离最小，可能是填充效果最好的一种。然而一般来说，该方法很费时，当数据规模很大、空值很多的时候，该方法是不可行的。
4. **特殊值填充**（Treating Missing Attribute values as Special values）将空值作为一种特殊的属性值来处理，它不同于其他的任何属性值。如所有的空值都用“unknown”填充。这样将形成另一个有趣的概念，可能导致严重的数据偏离，一般不推荐使用。
5. **平均值填充**（Mean/Mode Completer）将初始数据集中的属性分为数值属性和非数值属性来分别进行处理。 如果空值是数值型的，就根据该属性在其他所有对象的取值的平均值来填充该缺失的属性值；如果空值是非数值型的，就根据统计学中的众数原理，用该属性在其他所有对象的取值次数最多的值(即出现频率最高的值)来补齐该缺失的属性值。与其相似的另一种方法叫条件平均值填充法（Conditional Mean Completer）。在该方法中，用于求平均的值并不是从数据集的所有对象中取，而是从与该对象具有相同决策属性值的对象中取得。
6. **热卡填充**（Hot deck imputation，**或就近补齐**）对于一个包含空值的对象，热卡填充法在完整数据中找到一个与它最相似的对象，然后用这个相似对象的值来进行填充。不同的问题可能会选用不同的标准来对相似进行判定。该方法概念上很简单，且利用了数据间的关系来进行空值估计。这个方法的缺点在于难以定义相似标准，主观因素较多。
7. **K最近距离邻法（K-means clustering）**先根据欧式距离或相关分析来确定距离具有缺失数据样本最近的K个样本，将这K个值加权平均来估计该样本的缺失数据。使用所有可能的值填充，用空缺属性值的所有可能的属性取值来填充，能够得到较好的补齐效果。但是，当数据量很大或者遗漏的属性值较多时，其计算的代价很大，可能的测试方案很多。
8. **组合完整化方法（Combinatorial Completer）**

用空缺属性值的所有可能的属性取值来试，并从最终属性的约简结果中选择最好的一个作为填补的属性值。这是以约简为目的的数据补齐方法，能够得到好的约简结果；但是，当数据量很大或者遗漏的属性值较多时，其计算的代价很大。

1. **回归**（Regression）基于完整的数据集，建立回归方程。对于包含空值的对象，将已知属性值代入方程来估计未知属性值，以此估计值来进行填充。当变量不是线性相关时会导致有偏差的估计。
2. **期望值最大化方法（**Expectation maximization，EM）EM算法是一种在不完全数据情况下计算极大似然估计或者后验分布的迭代算法。在每一迭代循环过程中交替执行两个步骤：E步（Excepctaion step,期望步），在给定完全数据和前一次迭代所得到的参数估计的情况下计算完全数据对应的对数似然函数的条件期望；M步（Maximzation step，极大化步），用极大化对数似然函数以确定参数的值，并用于下步的迭代。算法在E步和M步之间不断迭代直至收敛，即两次迭代之间的参数变化小于一个预先给定的阈值时结束。该方法可能会陷入局部极值，收敛速度也不是很快，并且计算很复杂。
3. **不处理**，补齐处理只是将未知值补以我们的主观估计值，不一定完全符合客观事实，在对不完备信息进行补齐处理的同时，我们或多或少地改变了原始的信息系统。而且，对空值不正确的填充往往将新的噪声引入数据中，使挖掘任务产生错误的结果。因此，在许多情况下，我们还是希望在保持原始信息不发生变化的前提下对信息系统进行处理。不处理缺失值，直接在包含空值的数据上进行数据挖掘的方法包括贝叶斯网络和人工神经网络等。

**总结：**根据电影票房分析的需求，准备使用平均值填充法进行数据补齐，在指标量化的过程中参考了很多数据，这些数据可能来自不同的数据库，不可避免的会造成一些值的空缺，所以有些数据设置成可为空值。具体问题具体分析，我们可以根据上面介绍的空缺值处理方法进行数据的处理。

1. 模型建立目标（量化后的数据表）

表1-1 变量说明

|  |  |  |  |
| --- | --- | --- | --- |
| **变量名称** | **变量说明** | **变量来源** | **变量属性** |
| 总票房 Y | 电影票房的总收入（单位万元） | 中国票房、豆瓣数据库 | 因变量 |
| 导演 X1 | 量化后导演影响力 | 豆瓣、百度、微博 | 自变量  量化变量 |
| 演员 X2 | 量化演员的影响力 |
| 编剧 X3 | 量化编剧的影响力 |
| 制作公司 X4 | 制作公司影响力 | 艺恩电影数据库  豆瓣电影数据库 |
| 发行公司 X5 | 发行公司影响力 |
| 类型 X6 | 不同类型量化值 |
| 档期 X7 | 不同档期量化值 |  |
| 知识产权 X8 | IP影响力 |  |
| 制作技术 X9 | 制作技术影响力 |  |
| 排片数据 X10 | 排片的影响力 |  |
| 点映 X11 | 有设1无设0 | 历史来源未知 | 虚拟变量 |
| 网络口碑 X12 | 网络口碑影响力 | 数据来源未知 | 自变量 |

先建立线性回归模型（结果不好考虑其他建模方式），线性回归模型公式：



虚拟变量设置的原则，在模型中引入多个虚拟变量时，虚拟变量的个数应按下列原则确定：如果回归模型有截距项有m种互斥的属性类型，在模型中引入（m-1）个虚拟变量。如果回归模型无截距项，有m个特征，设置m个虚拟变量。电影的类型的细分设置成虚拟变量，使用0|1的值来表示该类型的有无（备注:15类会不会太多）、电影的制作技术细分为三个虚拟变量（IMAX,3D,2D），档期分类量化分析也可以设置虚拟变量。

**遇到的问题：**

1. 由于系统使用爬虫收集信息，所以数据来源虽然广泛，但是数据的一致性很差。例如两个数据库的储存相同类型值的方式不同甚至是值不同。或者从一个数据库收集到的信息，在另一个数据库中很难找到对应的值，或者说数据缺失严重。初步的想法是针对每一个问题编写程序进行数据一致性的处理（数据清洗）。缺失值严重时（大于20%时）考虑删除缺失严重的样本或者重新选择变量。
2. 空缺值详细的处理方法，现在还在数据收集阶段，没有实际数据的缺失情况，所以缺失值的处理方法也要随下一步的进行而制定。
3. 对于不同的指标的数据分析和量化是多个过程，每个过程需要单独进行或者综合几个过程一起进行数据分析。例如：我们已经获得了发行公司、制作公司及其对应历史电影的票房数据，根据这些数据也可以简单分析发行公司和制作公司的影响力。

参考文献：

[1] CSDN博客.数据缺失的处理方法