影视量化分析数据库设计

简而言之，这个数据库的总体设计思想很简单，即对每一个需要量化的自变量的相关信息建立一张数据表格。初步决定选用关系型数据库，并且以电影概要信息表（movie）为中心，分析每一张表对应的实体和电影实体之间的关系和各自的属性、并且对每一个指标都建立一张单独的数据表，最后建立一张量化后的数据表（包含所有自变量和因变量）。

* + 电影概要信息表（movie）：movie表主要储存电影的主要信息，如导演名、演员名、编剧名、类型、制作技术、制作和发行公司等。
  + 变量信息的数据表：一共有12个变量对应了11张表（注：点映直接当做电影属性，取值布尔型0|1）。
  + 量化后的数据表（movie\_quantitied）：movie\_quantitied主要保存各个变量量化后的数值。

1. **ER图样例**

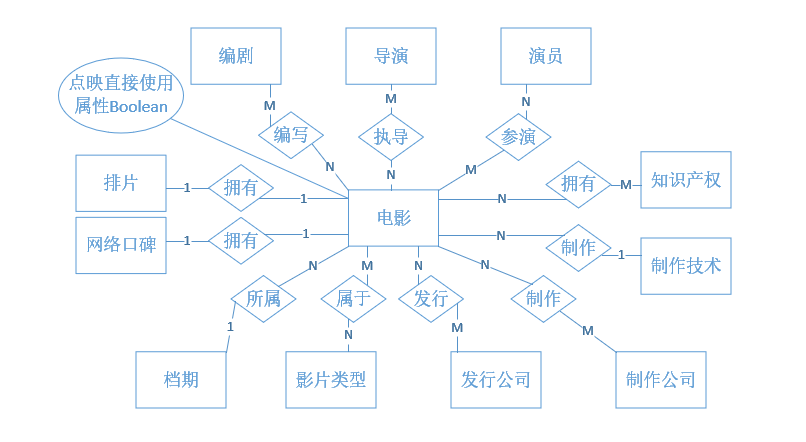


图3-1 系统总体ER图（注：属性太多，下面详细写）

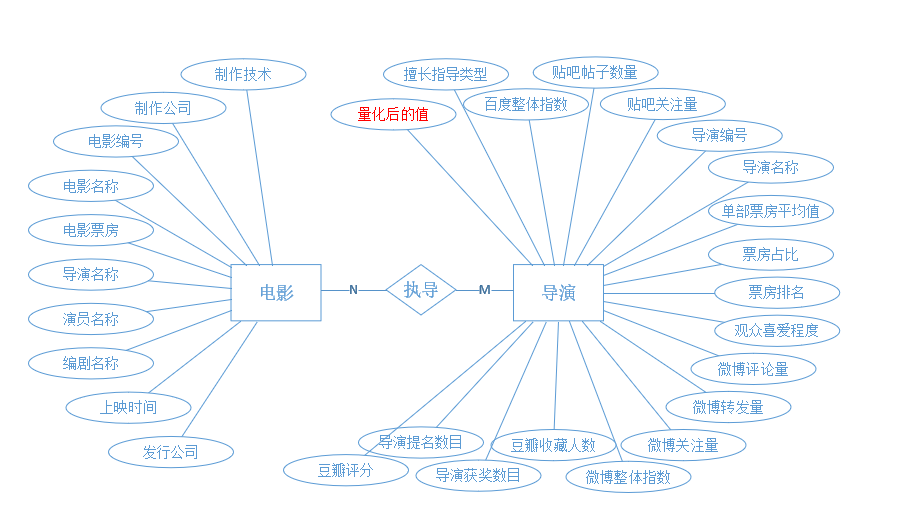


图3-2 导演的ER图

根据ER图转化并设计概要的数据表：

导演表（导演编号，导演名称，…）

执导表（执导编号，导演编号，电影编号）

电影表（电影编号，电影名称，…）

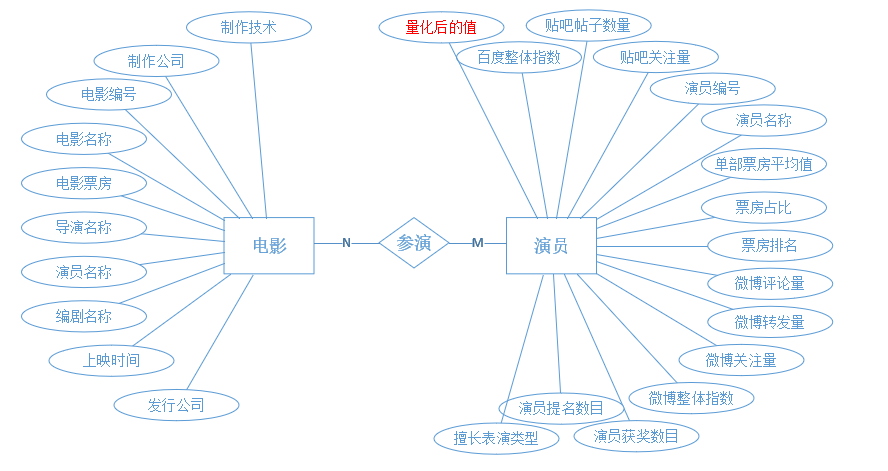


图3-3 演员的ER图

根据ER图转化并设计概要的数据表：

演员表（演员编号，演员名称，…）

参演表（参演编号，演员编号，电影编号）（附加表）

电影表（电影编号，电影名称，…）

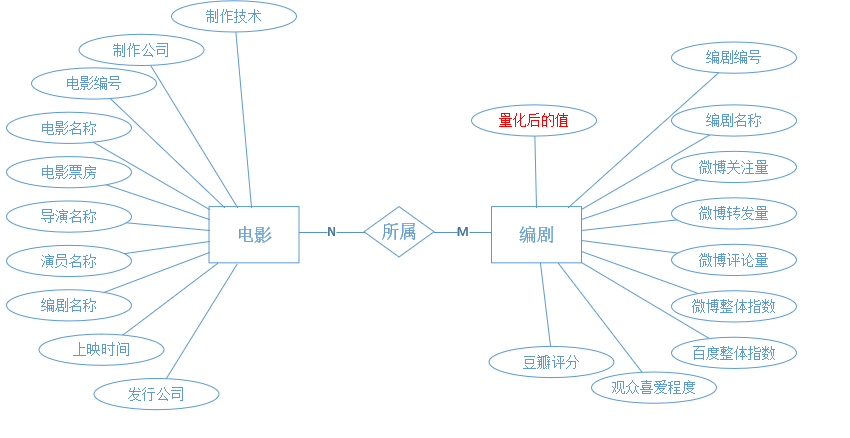


图3-4 编剧的ER图

根据ER图转化并设计概要的数据表：

编剧表（演员编号，演员名称，…）

编写表（编写编号，编剧编号，电影编号，剧本名）（附加表）

电影表（电影编号，电影名称，…）

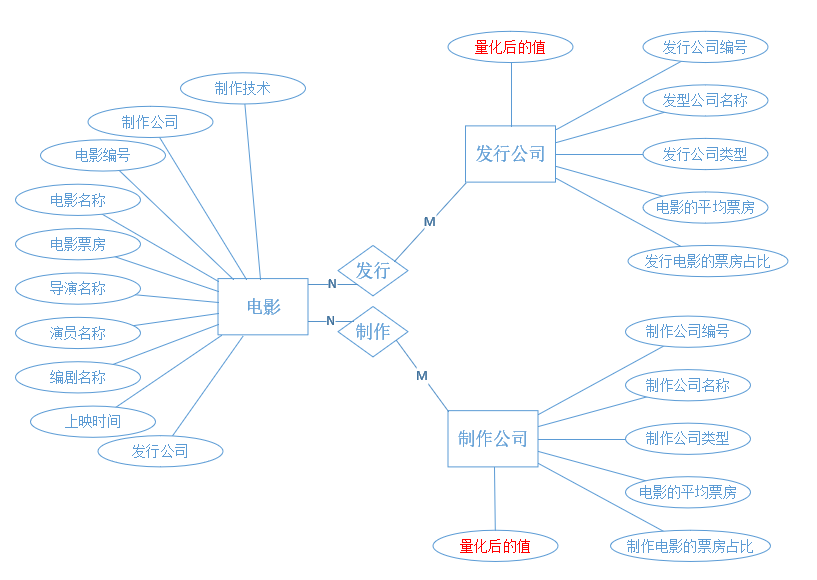


图3-5 发行公司和制作公司ER图

根据ER图转化并设计概要的数据表：

制作公司表（制作公司编号，制作公司名称，…）

制作表（制作编号，制作公司编号，电影编号）（附加表）

电影表（电影编号，电影名称，…）

发行公司表（发行公司编号，发行公司名称，…）

发行表（发行编号，发行公司编号，电影编号）（附加表）

电影表（电影编号，电影名称，…）

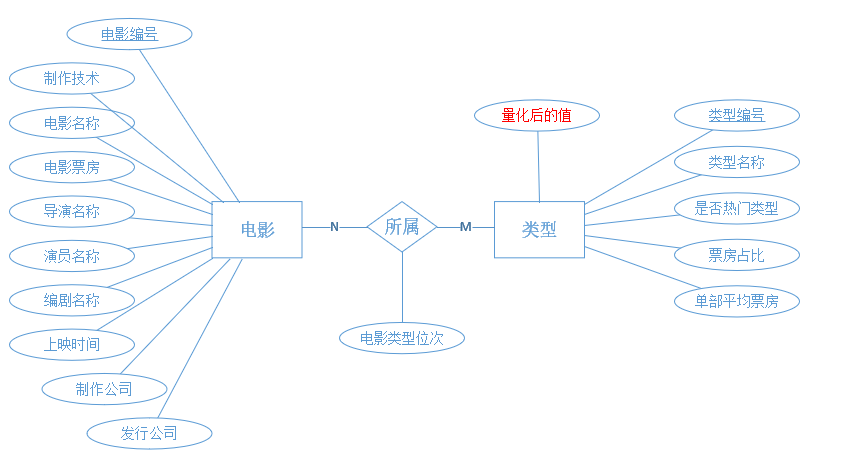


图3-5 影片类型ER图

根据ER图转化并设计概要的数据表：

类型表（类型编号，类型名称，…）

所属表（所属编号，类型编号，电影编号，电影类型的位次）（附加表）

电影表（电影编号，电影名称，…）

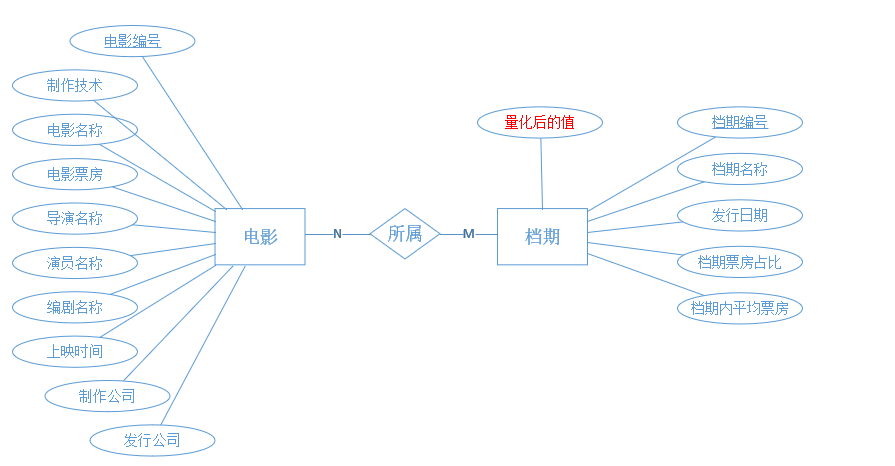


图3-6 影片档期ER图

根据ER图转化并设计概要的数据表：

档期表（档期编号，档期名称，…）

所属表（所属编号，档期编号，电影编号）（附加表）

电影表（电影编号，电影名称，…）

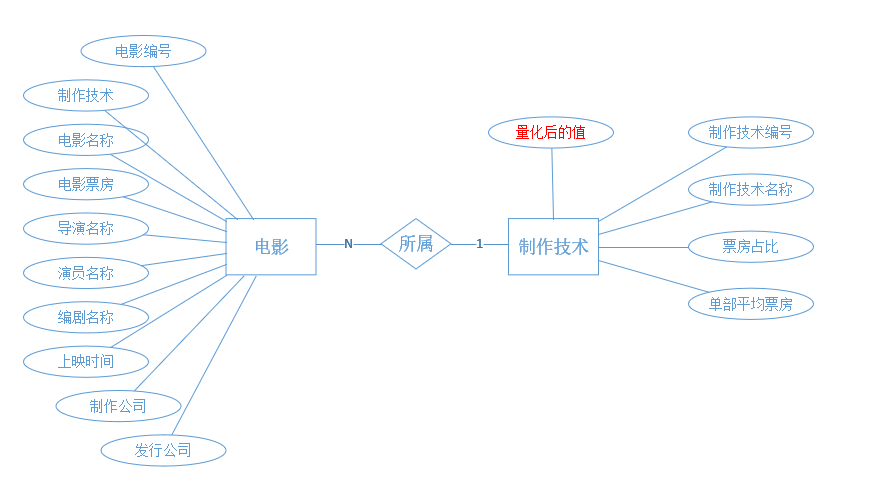


图3-7 制作技术ER图

根据ER图转化并设计概要的数据表：（制作技术编号0-2，只有三类制作技术）

制作技术表（制作技术编号，制作技术名称，…）

电影表（电影编号，电影名称，制作技术编号，…）

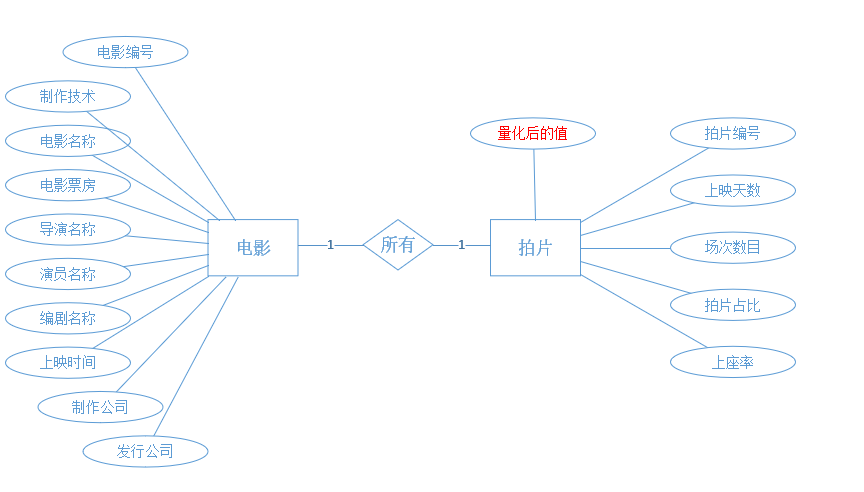


图3-8 拍片的影响力

根据ER图转化并设计概要的数据表：

拍片表（拍片编号，上映天数，场次数目，拍片占比，上座率）

电影表（电影编号，电影名称，拍片编号，…）

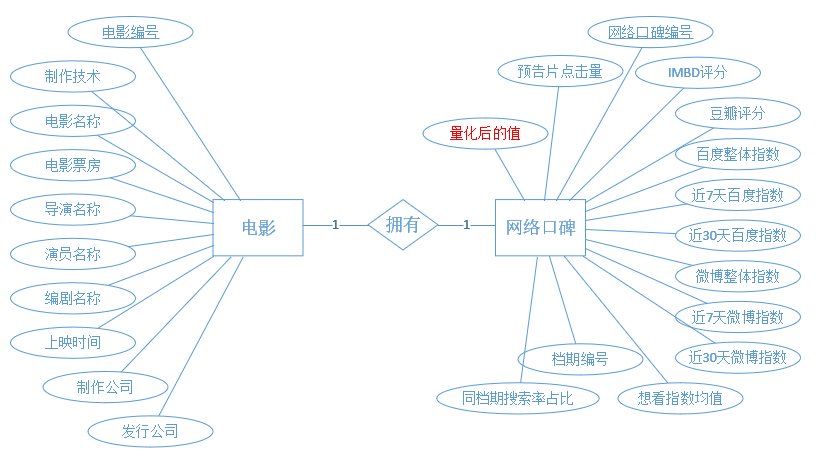


图3-9 网络口碑ER图

根据ER图转化并设计概要的数据表：

网络口碑表（网络口碑编号，豆瓣评分，…）

电影表（电影编号，电影名称，网络口碑编号，…）

1. 建立详细的数据表（先英文化）

根据现在收集到的数据（3250余部国产电影的样本），首先选用MySQL数据库建立了一个待分析的电影概要信息的数据表（movie），在此之后的所有表格的信息都可以根据**电影概要数据表(movie)**来动态获取。我们准备把收集到的每一个指标的数据都单独存入相应的数据表中，然后使用程序量化分析后动态生成一个量化后的数据表，整个数据库的数据表为：1（概要数据表）+11（详细的指标数据表）+1（量化后数据表） = 13个独立的数据表，同时要保证量化数据表和指标数据表的数据一致性。

表3-1 电影概要信息数据表（movie）

|  |  |  |  |
| --- | --- | --- | --- |
| 指标名称 | 数据表中属性名称 | 属性值类型 | 备注信息 |
| 电影ID | movie\_id | INT(10) | 主键 |
| 电影名称 | movie\_name | VARCHAR（45） | 电影名 |
| 导演名称 | director\_name | VARCHAR（45） | 导演名 |
| 演员列表 | actors\_name | TEXT（500） | 演员有排名 |
| 编剧名称 | screenwriter\_name | VARCHAR（45） | 编剧 |
| 上映时间 | issue\_date | DATE | 转化成档期 |
| 票房总量 | box\_office | MONEY | 作为分析参照 |
| 电影类型 | movie\_type | VARCHAR（45） | 多类型掺杂 |
| 制作技术 | production\_technology | VARCHAR（45） | 2D,3D,IMAX |
| 制作公司列表 | production\_company | TEXT（500） | 制作公司 |
| 发行公司列表 | distribution\_company | TEXT（500） | 发行公司 |

我们准备根据表1所示的数据表来动态生成量化后的数据表2，主要是根据表1中的概要信息，在互联网收集数据并设计量化方案，以此为依据来动态建立量化后的数据表2。

表3-2 量化后的数据表（quantified \_data）

|  |  |  |  |
| --- | --- | --- | --- |
| 影响因素名称 | 数据表中属性名称 | 属性值类型 | 备注信息 |
| 影响因素ID | factors\_id | INT（自增长） | 主键 |
| 电影ID | movie\_id | INT（10） | 电影ID |
| 导演的影响力 | movie\_director | INT（10） | 1-100 |
| 演员的影响力 | movie\_actors | INT（10） | 1-100 |
| 编剧的影响力 | issue\_time | INT（10） | 1-100 |
| 制作公司影响力 | production\_company | INT（10） | 1-100 |
| 发行公司影响力 | distribution\_company | INT（10） | 1-100 |
| 影片类型影响力 | movie\_type | INT（10） | 15种电影类型根据不同位次加权  1-100 |
| 制作技术影响力 | production\_technology | INT（10） | 3种制作技术  1-100 |
| 档期影响影响力 | issue\_date | INT（10） | 6种档期  1-100 |
| 知识产权影响力 | intellectual\_property | INT（10） | 1-100 |
| 排片影响力 | row\_of\_tablets | INT（10） | 1-100 |
| 搜索指数影响力 | search\_index | INT（10） | 1-100 |
| 网络口碑影响力 | reputation | INT（10） | 1-100 |
| 点映影响力 | on\_demand | BOOLEAN | 1或0 |
| 票房 | box\_office | INT（10） | 万元人民币为单位 |

量化的数据表来自于每个详细指标的数据表，详细指标的数据表在收集好数据后建立，还有许多可能需要更改的地方。

**遇到的问题：**

1. 样本多了不可能人工改写数据，不同值的处理方式不同，为了保证建立好整个数据库，数据处理（数据清洗）的工作量会很大，考虑使用多种方式进行数据处理。