Sy6-图书馆借阅书籍推荐分析

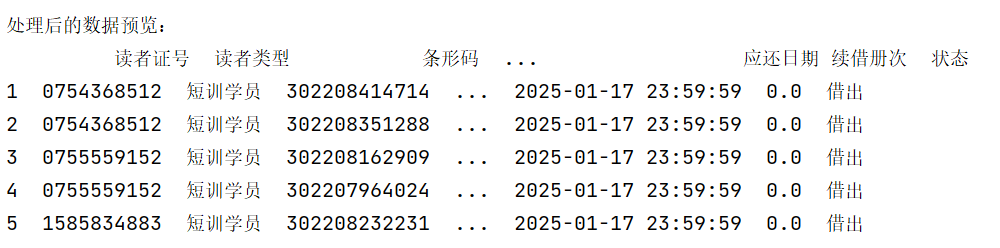
1. 生成订单数据集

**1.数据清洗**

原始数据集包含5865条记录，13个字段，数据内容包括读者证号、图书条形码、题名、责任者、出版社等。首先修正题名列中非字符串值，删除空值和无效记录，删除重复记录，确保数据质量。

**2. 生成订单数据集**

根据每位读者的借阅记录，合并生成订单数据集（即每位读者对应一个包含所有借阅图书的列表），并将数据保存至新文件book\_orders.csv中，处理后的数据如下图所示：



1. 利用关联分析方法实现图书之间关联关系的分析
2. **数据过滤**

为提升分析效率，在此过滤掉低频图书和不完整订单，统计图书的借阅频率，选择借阅频率最高的500本书，并删除借阅图书少于2本的订单，共删除1829条记录。最终保留228名读者、362本图书。

1. **构建二元矩阵**

将每本书作为矩阵的列，矩阵的行对应读者，读者借阅行为（是否借阅某本书）作为矩阵的值（0或1），构建出二元矩阵，若某位读者借阅了某本书，则对应单元格的值为1，否则为0。

1. **生成频繁项集和关联规则**

使用Apriori算法对二元矩阵进行分析，提取频繁项集。实验可根据数据集大小（读者数量）动态调整支持度阈值，确保分析的有效性。本实验中，支持度阈值设定为0.0132，即频繁项集至少需要被3名读者借阅。

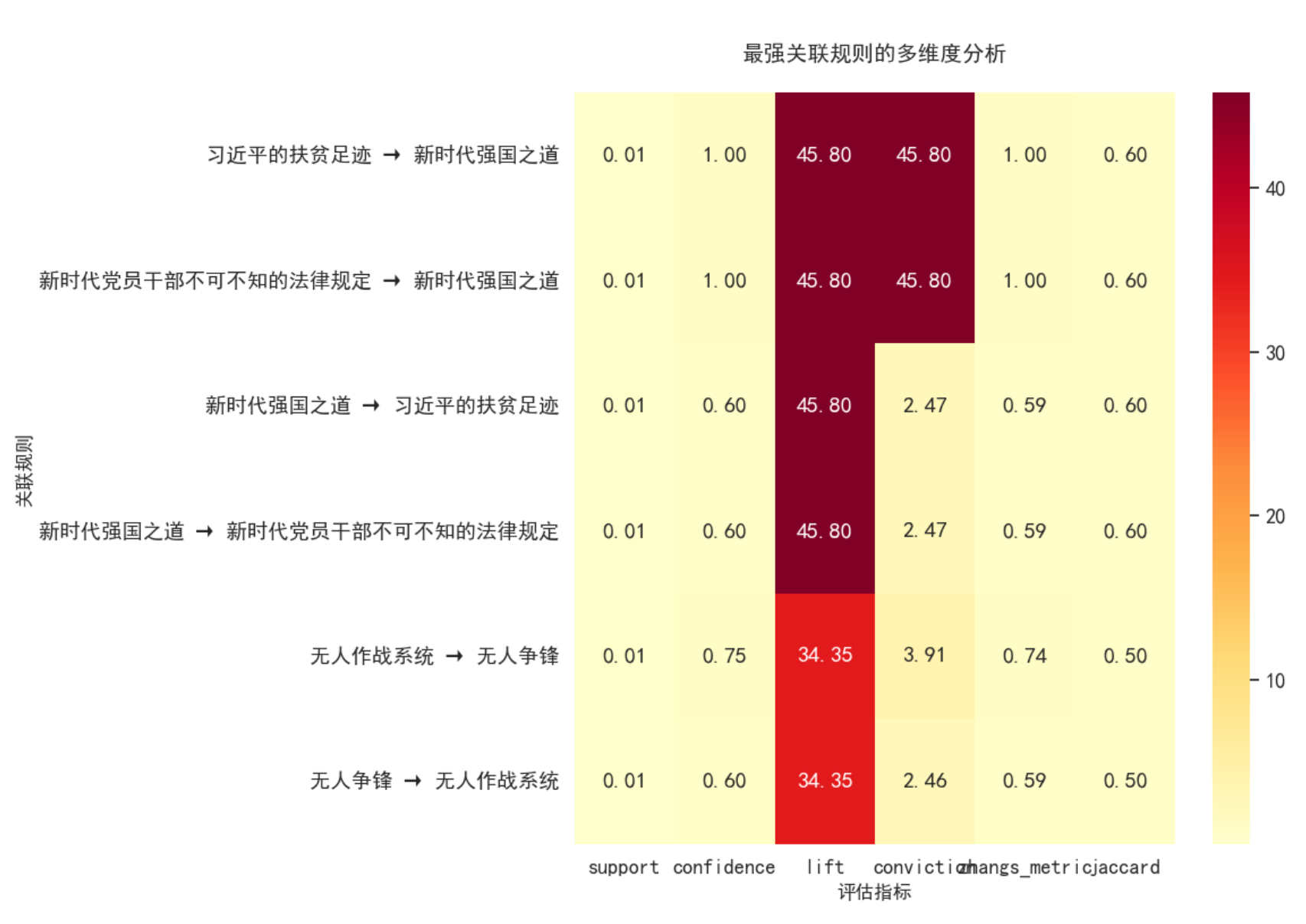
在生成频繁项集后，进一步生成关联规则，计算每条规则的质量指标，包括支持度、置信度和提升度。

其中，支持度 = 规则出现次数 / 总记录数，表示规则中的所有图书同时被借阅的概率。

置信度 = 同时借阅A和B的次数 / 借阅A的次数，表示在借阅了前项图书的情况下，同时借阅后项图书的条件概率。

提升度 = 置信度 / 后项的支持度，表示前项和后项之间的关联强度，值越大表明关联性越强。

筛选提升度大于1的规则，共生成6条有效规则。最终得到的强关联关系如下图所示



1. 结论



从结果中可以看出，提升度最高的规则为习近平的扶贫足迹 → 新时代强国之道，提升度为45.6，表明借阅这两本书的关联性极强。即如果读者借阅了《习近平的扶贫足迹》，他很可能同时会借阅《新时代强国之道》。且政治、经济类书籍之间的关联性强，符合读者群体的借阅习惯。

实验得到的强关联规则可用于指导图书推荐和优化书架布局，为图书馆管理提供数据支持。