華中科技大学

课 程 实 验 报 告

课程名称： 大数据分析

专业班级： **CS2202**

学　　号： **U202215399**

姓　　名： 吴渝东

指导教师： 崔金华

报告日期： **2024.06.20**

计算机科学与技术学院

目录

实验三 关系挖掘实验.....1
3.1 实验内容.....1
3.2 实验过程.....1
3.2.1 编程思路.....1
3.2.2 遇到的问题及解决方式.....2
3.2.3 实验测试与结果分析.....2
3.3 实验总结.....4

实验三 关系挖掘实验

3.1 实验内容

必做：

1. 实验内容

实验数据：从实验一中得到的引用关系数据为<<title,<title1,…,titlek>>,…>，将其处理为<<title,title1,…,titlek>,…>作为算法输入。

编程实现 Apriori 算法，要求使用实验一得到的前 1000 个 title 及其引用关系作为实验数据。

2. 实验要求

输出 1~4 阶频繁项集与关联规则，各个频繁项的支持度，各个规则的置信度，各阶频繁项集的数量以及关联规则的总数。

固定参数以方便检查，频繁项集的最小支持度为 0.15，关联规则的最小置信度为 0.3。此处支持度的定义为某个项集出现的频率，也就是包含该项集的数目与总数目的比例（总的购物篮数目为 1000）。

加分项：

1. 实验内容

在 Apriori 算法的基础上，使用 pcy 算法对二阶频繁项集的计算阶段进行优化。

2. 实验要求

输出 1~4 阶频繁项集与关联规则，各个频繁项的支持度，各个规则的置信度，各阶频繁项集的数量以及关联规则的总数。

输出 pcy 算法中的 vector 的值，以 bit 位的形式输出。

固定参数以方便检查，频繁项集的最小支持度为 0.15，关联规则的最小置信度为 0.3。

3.2 实验过程

3.2.1 编程思路

首先，根据购物篮模型，将实验一获得的跳转关系视为购物篮的集合，同时，将实验一 reduce 得到的单词视为所有商品的集合。

计算 1 阶频繁项集：将所有项的集合作为候选项集，遍历候选项集中的元素，计算支持度，大于最小支持度的则为 1 阶频繁项集。

计算 n 阶频繁项集：首先，利用 1 阶频繁项集和 n-1 阶频繁项集计算候选项集，遍历两个频繁项集中的元素 elem1,elem2,如果 elem1 不是 elem2 的子集，合并两个集合得到候选项集。然后对候选项集进行遍历，计算其支持度，大于最小支持度的则为 1 阶频繁项集。

PCY 算法：以上述同样的方式获取候选项集，然后遍历候选项集中的元素 elem，计算 elem 的哈希值，根据哈希值哈希到对应的桶中，elem 每在购物篮中出现一次，对应的哈希桶计数值加 1，然后再次遍历候选项集中的元素，选出其中哈希到频繁桶的元素，计算其支持度，判断是否频繁项集。

计算关联规则：将所有频繁项集及其对应的支持度作为输入，对于 n 阶频繁项集，首先获取其所有子集，然后遍历其子集，通过计算（子集支持度/本身支持度）得到该规则的置信度，大于最小置信度的则为我们感兴趣的关联规则。

3.2.2 遇到的问题及解决方式

问题：计算高阶频繁项集时，频繁项集的数量大于应有的数量

解决方式：通过分析生成的频繁项集发现，出现了重复的频繁项集，因此进一步分析发现，在生成候选项集时，简单将两个子集合并，没有去除其中的重复元素，因此，采用集合的数据类型，在生成候选项集时，自动过滤掉重复出现的项集。

3.2.3 实验测试与结果分析

实验运行得到的频繁项集和关联规则都如下图所示（内容过多，仅展示部分内容），程序正确实现了所需的功能。

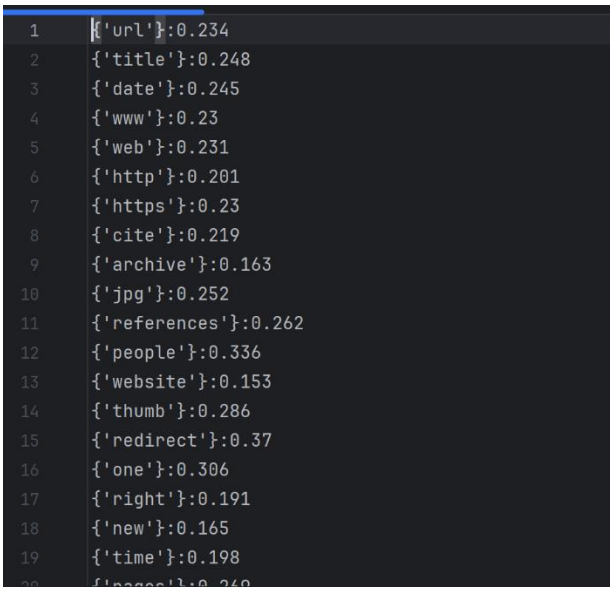

图 3-1 1 阶频繁项集

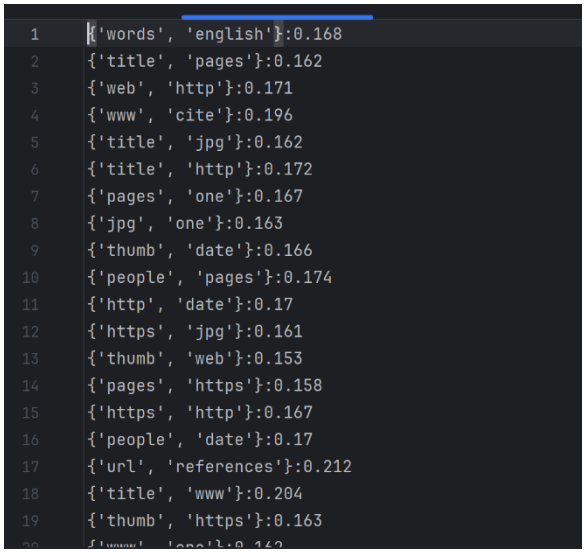

图 3-2 2 阶频繁项集

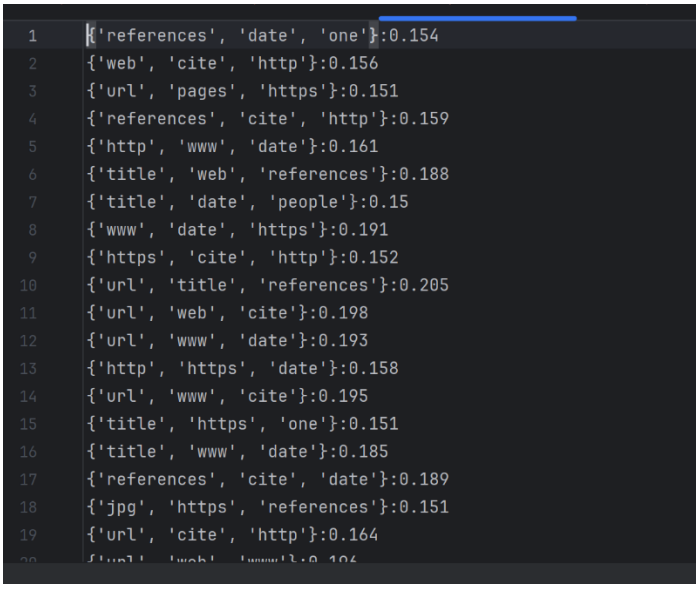

图 3-3 3 阶频繁项集

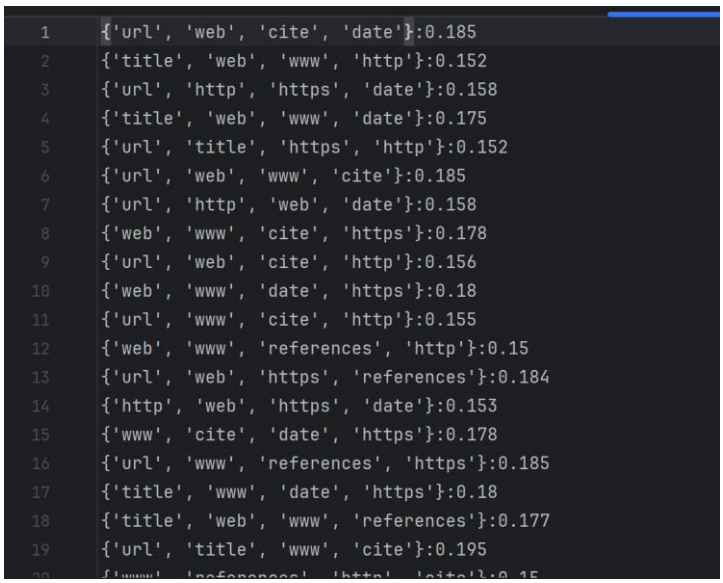

图 3-4 4 阶频繁项集

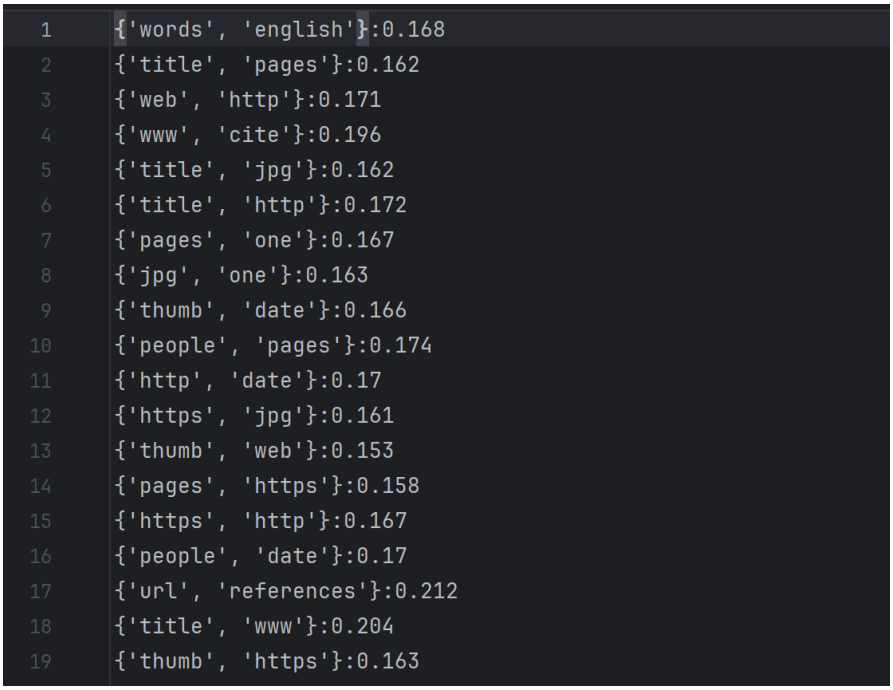

图 3-5 pcy 算法生成的 2 阶频繁项集

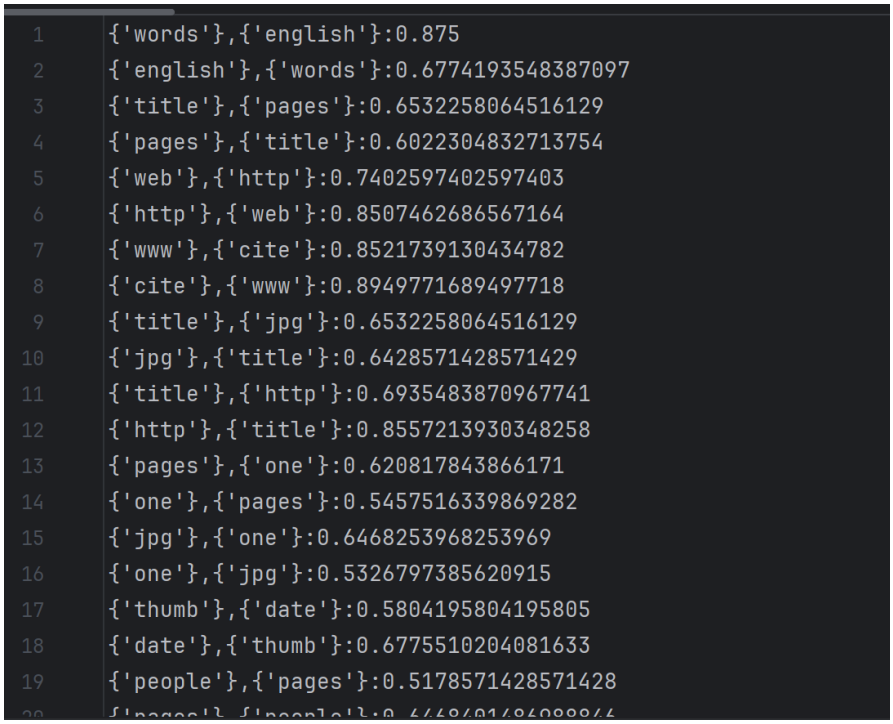

图 3-6 关联规则

3.3 实验总结

通过本次实验，深入了解了如何计算频繁项集，以及 pcy 算法的运行原理，对关联规则的生成也有了更加深刻的了解。同时本次实验让我体会到了，如何利用已有的内容对程序进行优化，例如计算高阶频繁项集时，无需对所有

的集合进行遍历，而是可以利用已有的低阶频繁项集来构建候选项集，从而减少需要遍历的元素数量，大大减少了工作量。