

华中科技大学

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 $\langle\langle\text{title}, \langle\text{title}_1, \dots, \text{title}_k\rangle\rangle, \dots\rangle$ ，将其处理为 $\langle\langle\text{title}, \text{title}_1, \dots, \text{title}_k\rangle, \dots\rangle$ 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。

```
1  {'url':0.234
2  {'title':0.248
3  {'date':0.245
4  {'www':0.23
5  {'web':0.231
6  {'http':0.201
7  {'https':0.23
8  {'cite':0.219
9  {'archive':0.163
10 {'jpg':0.252
11 {'references':0.262
12 {'people':0.336
13 {'website':0.153
14 {'thumb':0.286
15 {'redirect':0.37
16 {'one':0.306
17 {'right':0.191
18 {'new':0.165
19 {'time':0.198
20 {'pages':0.260
```

图 3-1 1 阶频繁项集

```

1  {'words', 'english'}:0.168
2  {'title', 'pages'}:0.162
3  {'web', 'http'}:0.171
4  {'www', 'cite'}:0.196
5  {'title', 'jpg'}:0.162
6  {'title', 'http'}:0.172
7  {'pages', 'one'}:0.167
8  {'jpg', 'one'}:0.163
9  {'thumb', 'date'}:0.166
10 {'people', 'pages'}:0.174
11 {'http', 'date'}:0.17
12 {'https', 'jpg'}:0.161
13 {'thumb', 'web'}:0.153
14 {'pages', 'https'}:0.158
15 {'https', 'http'}:0.167
16 {'people', 'date'}:0.17
17 {'url', 'references'}:0.212
18 {'title', 'www'}:0.204
19 {'thumb', 'https'}:0.163
20 {'thumb', 'date'}:0.166

```

图 3-2 2 阶频繁项集

```

1  {'references', 'date', 'one'}:0.154
2  {'web', 'cite', 'http'}:0.156
3  {'url', 'pages', 'https'}:0.151
4  {'references', 'cite', 'http'}:0.159
5  {'http', 'www', 'date'}:0.161
6  {'title', 'web', 'references'}:0.188
7  {'title', 'date', 'people'}:0.15
8  {'www', 'date', 'https'}:0.191
9  {'https', 'cite', 'http'}:0.152
10 {'url', 'title', 'references'}:0.205
11 {'url', 'web', 'cite'}:0.198
12 {'url', 'www', 'date'}:0.193
13 {'http', 'https', 'date'}:0.158
14 {'url', 'www', 'cite'}:0.195
15 {'title', 'https', 'one'}:0.151
16 {'title', 'www', 'date'}:0.185
17 {'references', 'cite', 'date'}:0.189
18 {'jpg', 'https', 'references'}:0.151
19 {'url', 'cite', 'http'}:0.164
20 {'url', 'web', 'www'}:0.194

```

图 3-3 3 阶频繁项集

```

1  {'url', 'web', 'cite', 'date'}:0.185
2  {'title', 'web', 'www', 'http'}:0.152
3  {'url', 'http', 'https', 'date'}:0.158
4  {'title', 'web', 'www', 'date'}:0.175
5  {'url', 'title', 'https', 'http'}:0.152
6  {'url', 'web', 'www', 'cite'}:0.185
7  {'url', 'http', 'web', 'date'}:0.158
8  {'web', 'www', 'cite', 'https'}:0.178
9  {'url', 'web', 'cite', 'http'}:0.156
10 {'web', 'www', 'date', 'https'}:0.18
11 {'url', 'www', 'cite', 'http'}:0.155
12 {'web', 'www', 'references', 'http'}:0.15
13 {'url', 'web', 'https', 'references'}:0.184
14 {'http', 'web', 'https', 'date'}:0.153
15 {'www', 'cite', 'date', 'https'}:0.178
16 {'url', 'www', 'references', 'https'}:0.185
17 {'title', 'www', 'date', 'https'}:0.18
18 {'title', 'web', 'www', 'references'}:0.177
19 {'url', 'title', 'www', 'cite'}:0.195
20 {'www', 'references', 'http', 'cite'}:0.15

```

图 3-4 4 阶频繁项集

1	{'words', 'english':0.168
2	{'title', 'pages':0.162
3	{'web', 'http':0.171
4	{'www', 'cite':0.196
5	{'title', 'jpg':0.162
6	{'title', 'http':0.172
7	{'pages', 'one':0.167
8	{'jpg', 'one':0.163
9	{'thumb', 'date':0.166
10	{'people', 'pages':0.174
11	{'http', 'date':0.17
12	{'https', 'jpg':0.161
13	{'thumb', 'web':0.153
14	{'pages', 'https':0.158
15	{'https', 'http':0.167
16	{'people', 'date':0.17
17	{'url', 'references':0.212
18	{'title', 'www':0.204
19	{'thumb', 'https':0.163

图 3-5 pcy 算法生成的 2 阶频繁项集

1	{'words'},{'english':0.875
2	{'english'},{'words':0.6774193548387097
3	{'title'},{'pages':0.6532258064516129
4	{'pages'},{'title':0.6022304832713754
5	{'web'},{'http':0.7402597402597403
6	{'http'},{'web':0.8507462686567164
7	{'www'},{'cite':0.8521739130434782
8	{'cite'},{'www':0.8949771689497718
9	{'title'},{'jpg':0.6532258064516129
10	{'jpg'},{'title':0.6428571428571429
11	{'title'},{'http':0.6935483870967741
12	{'http'},{'title':0.8557213930348258
13	{'pages'},{'one':0.620817843866171
14	{'one'},{'pages':0.5457516339869282
15	{'jpg'},{'one':0.6468253968253969
16	{'one'},{'jpg':0.5326797385620915
17	{'thumb'},{'date':0.5804195804195805
18	{'date'},{'thumb':0.6775510204081633
19	{'people'},{'pages':0.5178571428571428
20	{'pages'},{'people':0.4449701794088864

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。