2019/02/09 读书报告

学号 71117228 姓名 李泓烨

读书进度：2.5

一、读书报告内容

1. 自己提出问题的理解

① 为什么在 Level2-candidate-gen()函数中，第五行 sup(l)不会比 sup(h)小？

讨论后的理解：这里只是说 sup(l)可能比 sup(h)少，并不一定是绝对的，这里我看的是英文版的书，翻译的时候理解有误

② 怎么理解 MScandidate-gen()函数中的 pruning step？为什么会存在那个 exception？

讨论后的理解：对于 c 的每个(k-1)-子集子集 s，如果 s 不在F_{k-1}中，且 s 包含 c[1]，**也即 c 中拥有最低 MIS 值的项目**，则 c 就能从C_k中删除。而如果这个 s 不包含 c[1]，即使 s 并不在F_{k-1}中，我们也不能把 c 删除, 因为我们不能断定 s 会否满足 MIS(c[1]), 尽管我们知道 s 不满足 MIS(c[2]), 除非 MIS(c[1]) = MIS(c[2])(算法第 9 行)

2. 别人提出问题的理解

① 提出的问题 1 :为什么在 MScandidate 函数中每次循环中使$|\sup(i_{k-1}) - \sup(i'_{k-1})| \leq \phi$ 就能令一个 itemset 中 maxsup 与 minsup 的差值在 ϕ 的范畴内?求严谨数学说明。

自己的理解：对于已经生成的(k-1)-频繁项目集 f1 和 f2，其中的任意两项 x，y 必然满足|sup(x) − sup(y)| ≤ ϕ，如果要由 x，y 生成C_k，则唯一可能不满足支持度差别限制的两个项目就是i_{k-1}和i'_{k-1}

② 提出的问题 2：算法中为什么记录了 f − {a}就能确保所需的非 frequent 的 condition 的 count 被记录下来？

自己的理解 :因为迭代生成, 对于每一个频繁项目集减去其中 MIS 值最小的 count 都统计过了，这样在生成规则的时候，对于前件非频繁项目集有记录

③ 提出的问题 3：为什么对于使用 MIS 的 item 来说，向下闭包性质不再使用

自己的理解：如果一个项目集是频繁的，则根据 MS-Apriori 算法，条件之一是需要其真实的 support 满足集合中项目的最小 MIS，也就是要真实 support 大于第一个项目的 MIS。但是存在这样一种情况：频繁项目集F_k一个项目 f1 有一个(k-1)-子集 f2，f2 不包含 f1 中的第一个项目，并且 MIS(c[2]) > MIS(c[1])，则这个子集的真实 support 可能小于它的所有项目的 MIS，即 f2 可能是一个非频繁项目集，此时向下闭包性质不再适用。书上的反例可以帮助我们理解。

④ 提出的问题 4：在 MS-Apriori 算法中，level2-candidate-gen(L, ϕ)和 MScandidate-gen(Fk−1, ϕ）的内容有类似的地方，也有不同的地方，为什么两种 function 的内容有所不同。

自己的理解：因为 level2-candidate-gen 函数负责生成C_2，只需要考虑要生成的含有两项的项目集中，是否两个项目的真实支持度都满足第一个项目 MIS，这样既可保证满足规则的最小支持度要求；而 MScandidate-gen 中，因为有之前提到的 exception 的存在，需要多加一个判断 c[1]是否在自己中和 MIS(c[1])和 MIS(c[2])是否相等的操作。

⑤ 提出的问题 5：如何理解 Fig2.8 的 9-11 行？
自己的理解：这个问题和我自己提出的问题 2 是相同的，问的是对剪枝过程的理解，在此不再赘述。

⑥ 提出的问题 6：MS-Apriori 算法中，第一次为什么要产生 F1 和 L 两个集合？
自己的理解：因为 k=2 的时候是一个特例，C2 不能由 F1 直接产生，因为根据 exception，candidate 集合并不能直接由上一步得到的频繁集直接生成，否则不能包含全部的 candidate 项目集。

⑦ 提出的问题 7：对于 table data set 的 join step 如何在原先生成 candidate 算法的基础上进行调整，使得满足不会产生一个 candidate itemset containing two items from the same attribute,这里是什么意思，要如何调整(P22)
自己的理解：因为每个 item 中包含了 attribute 和 value，应该尽量避免产生候选项集中存在两个项，是同一 attribute 但具有不同的 value 值。虽然对于(attribute, value)，不同的 value 就已经对应不同的 pair，但是具有同一属性的两个项目出现在一个候选集合中是不合理的。

⑧ 提出的问题 8：反证法证明 2.4.3 的 lemma 的时候，为什么 a 属于 X 那么 X 就一定是一个 frequent itemset。
自己的理解：因为 a 属于 X，且 X 属于 f，a 的 MIS 值必然为 X 中最低的，相当于 X 的支持度也必然大于 MIS(a)

⑨ 提出的问题 9：当超市想要 X 和 Y 的关系时，需要同时算出 x->y 和 y->x 的 support 和 confidence 吗?
（因为它们的 confidence 不同，所以 x->y 和 y->x 不同？还是计算一个 support，计算两个 confidence 呢？）
自己的理解：看需求！

3. 读书计划
① 本周完成的内容章节：2.3-2.5
② 下周计划：看完第二章，第三章看到 3.2 左右