作业 4

朱志儒 SA20225085

6.2

（a）仅使用 G 和他们的支持度计数不能确定项集 A 是否频繁，需要闭频繁项集的信息。

算法：

输入：数据集 D 上所有闭频繁项集的集合 G，每个闭频繁项集 G[i]的支持度计数 C[i]，支持度阈值 e，给定项集 X；

输出：项集 X 是否为频繁的，如果是，给出 X 的支持度；

过程：对 G 中的闭频繁项集按支持度递增的次序排序，得到一个有序的闭频繁项集的集合G'，用$G'[i], C'[i]$分别表示第 i 个闭频繁项集和他的支持度计数

$for\ i = 1\ to\ |G'|$:

$if\ X \subset G'[i]$:

统计$X \subset G'[i]$中的支持度N_X，G'中所有集合的支持度计为N_G

$return\ \frac{N_X}{N_G}$

$return$ 项集 X 不是频繁的

（b）项集 X 是数据集 D 上的闭项集，如果不存在真超项集 Y 使得 Y 与 X 在 D 中具有相同的支持度计数。

项集 X 是数据集 D 上的生成元，如果不存在真子集$Y \subset X$使得$support(X) = support(Y)$。

显然，闭项集考虑的是真超项集，而生成元考虑的是真子集。闭频繁项集包含对应频繁项集的完整支持度信息，而频繁生成元不包含对应频繁项集的完整支持度信息。

6.6

C_1：项集	支持度计数
{M}	3
{O}	4
{N}	2
{K}	5
{E}	4
{Y}	3
{D}	1
{A}	1
{U}	1
{C}	2
{I}	1

L_1：项集	支持度计数
{M}	3
{O}	4
{K}	5
{E}	4
{Y}	3

C_2：项集	支持度计数
{M,O}	1
{M,K}	3
{M,E}	2
{M,Y}	2
{O,K}	3
{O,E}	3
{O,Y}	2
{K,E}	4
{K,Y}	3
{E,Y}	2

L_2：项集	支持度计数
{M,K}	3
{O,K}	3
{O,E}	3
{K,E}	4
{K,Y}	3

C_3：项集	支持度计数
{O,K,E}	3
{K,E,Y}	2

L_3：项集	支持度计数
{O,K,E}	3

频繁项集：{M}，{O}，{K}，{E}，{Y}，{M,K}，{O,K}，{O,E}，{K,E}，{K,Y}，{O,K,E}