第二次读书报告

侯宇捷

1. 自己提出的问题：

1.提出的问题：为什么在MS-candidate函数中每次循环中使|sup(ik-1) − sup(i’ k−1)| ≤ ϕ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。

讨论后的理解：最初思考时忽视了给出的f1、f2本身是频繁项集，可知其中每个项的mis差值都在ϕ的区间内。因此，在新一次的循环中，只需比较两个项集间不共有的项的实际support值是否符合要求即可。

2.提出的问题：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？

讨论后的理解：当时思考的时候考虑的是是否f-a就能包括所有的非频繁项集，然而忽略了算法实现的过程中，其实记录的并不仅仅是频繁项集的count，还有一些非频繁项集的、时候被排除的一些情况的count，这些count包含了f-a之外的一些非频繁项集情况，因此只需要额外记录f-a即可。

1. 别人提出的问题：

1.问题：fig2.8的第五行ik-1<ik-1‘处是否为了避免重复讨论，以及他们的大小是按什么标准比较的？

自己的理解：讨论后的结果是大小标准按什么排并不重要。但讨论后我又想了一下，觉得还是应该用mis排列。如果按字典序来排，那么在做F3的时候，i2与i2‘就不会按mis来排列，那么可能合并之后mis大的放在了第二个位置，小的放在第三个位置，如果此时mis3与mis1相同，那在算法第9行条件就应该更改，改成=mis2或=mis3。此外，用单一模式的排列有助于利用最早data pass时产生的M，在后几轮循环中找出符合要求的f1与f2。

2.问题：在MS-Apriori算法中第二轮的搜索与其他轮次的搜索不同的是什么？

自己的理解：其中参数是L而非F1.若是第二轮也与其他轮一样用F1在同样的函数中运行，则会导致一些符合条件的情况被遗漏。

3.问题：为什么在设定了mis的情况下，向下封闭性质不再适用？

自己的理解：向下封闭性质之所以之前能使用，是由于support计算公式上，随着取子集，分母不变，分子增大，因而support增大，而所需的support始终为定值，因此成立。然而现在，所要求的support对于itemset与其子集来说，未必一致，要看第一项是否保留（或

存在多个相同的最小mis），因此不再成立。

1. 读书计划

1.本周完成的内容章节：2.3-2.5，3.1部分

2.下周计划：第三章要求的全部内容

四、读书摘要

记录几条读书时令我印象深刻的内容：

1. 离散化：将确切的数字划分成一个个数字区间。如果统计时都只是一个个数字，数字重复的概率极低，单记录数字很难反映出数字间的规律及趋势，而将数字离散化、分成一个个区间来看就解决了这一问题。
2. 在MS-apriori函数中除了直接计算F1（这很容易得出）外，还计算了一个L。凡是单个item的count/n的值大于所有item中最小的mis的项，皆被纳入L中。第一次读的时候没感觉出这有什么意义，然而往下看发现第一次level-2的candidate函数用到的并非F1，而是L。数学上可以证明如果使用F1的话，可能会遗漏部分同样频繁的项集，而用L的话恰到好处：不是由L中的项构成的二元项集一定不频繁，而再将条件拉高，则会造成遗漏。
3. 尽管向下封闭性质并不成立，但如果一个itemset的子itemset与原itemset有相同的最小mis的话，若原itemset频繁，则子itemset也一定频繁。
4. 可以通过将某些项的mis设定到大于1的程度来保证生成的itemset中一定包含某些项的一部分或不止包含某些项。
5. Head-item的证明：通过反证法证明了如果单凭已经记录好的frequent itemset的support count无法计算confidence，则说明rule的condition一定非频繁，且f全集中的最小mis项一定在consequent中。
6. 设定mis的两种方法：（1）设定一个恒定的常数x，x在0到1之间，将这个x乘上每个item的实际support来得到mis（注：若用这种方式设定mis则mis与实际support有关，其余皆无关）（2）将一些frequency相近的item组成一组，并将这一组的所有item附上相同的mis。

7．在car-apriori的情况下，rule generation并没有前面的一些性质，因此就是简单的计算每一种可能。当然，由于class的特殊性与这种情况下所有rule的consequent仅为一class，因此也不存在combination explosion的情况。