web data mining 第三次读书报告

报告人：71117415-喻泽弘

读书进度：完成第三章的阅读

- 问题与解答
    1. 如何理解Fig 3.4的算法？(我提出的)

       该算法是递归产生决策树的过程，递归结束的条件是产生了leaf node，而结束分为三种情况：

       - 数据集D中以及只含有唯一的类别c_j，这说明划分以及完成
       - $A=\emptyset$,此时说明动用了所有的attribute进行判断，但是仍然不能确定数据子集的唯一类别，则将该处标为数据集中数量最多的类的leaf node
       - $P_0-P_g<threshold$,这说明即使还有attribute可以进行分类，但是对集合的纯度提升没有达到要求，将该处标为数据集中数量最多的类的leaf node

       如果不是在递归终止的条件下，则计算得出最能提高这个集合纯度的属性，令它称为该处的decision node，并且对这个attribute划分得出每个子集去除当前attribute后递归调用这个函数，该算法是一个贪婪算法，每次都选择最能提升集合纯度的attribute进行选择，一旦做出选择便无法更改

    2. 使用information ratio产生的问题？

       例如我们对学生做决策，以学号为属性，因为每个人只有一个学号，所以学号属性将学生划分为单元素集，$entropy_{学号}(D)=0$;但是也正是如此，因为学号对每个人是唯一的，所以从预测的观点而言，测试集中不会再出现相同的学号，因此这种划分虽然公式定义上降低了熵，但是毫无意义，并不是我们所想要寻找的

- 下周计划安排：

  完成web data mining第四章的阅读