# Apriori算法

## 基本概念：

学习Apriori算法首先要了解几个基本概念：

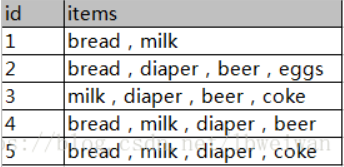
1. 项集：顾名思义，即项的集合。例如：牛奶、面包组成一个集合{牛奶、面包}，其中牛奶和面包为项，{牛奶、面包}为项集，称之为2项集。（说白了，其实就是集合）。
2. 支持度：项集A、B同时发生的概率称之为关联规则的支持度。如公式（1）

 （1）

1. 置信度：项集A发生的情况下，则项集B发生的概率为关联规则的置信度。如公式（2）

 （2）

支持度与置信度些许抽象，可以看一下举例说明：



**Fig1：概念举例说明**

如图数据为顾客购物情况，每一个id对应的items都是一个项集，现在需要对{milk，diaper}与{beer}关联性进行研究，计算支持度与置信度。

1. 计算支持度：计算{milk，diaper}{beer}同时发生的概率就相当于计算{milk，diaper，beer}出现的次数所占数据条的比重，即2/5.
2. 计算置信度：计算{milk，diaper}发生的情况下，则{beer}发生的概率就相当于计算{milk，diaper，beer}出现的次数所占{milk，diaper}发生次数的比重，即2/3.

最小支持度：最小支持度就是人为按照实际意义规定的阈值，表示项集在统计意义上的最低重要性。

最小置信度：最小置信度也是人为按照实际意义规定的阈值，表示关联规则最低可靠性。

频繁项集：满足最小支持度的所有项集，称作频繁项集。

（频繁项集性质：1、频繁项集的所有非空子集也为频繁项集；2、若A项集不是频繁项集，则其他项集或事务与A项集的并集也不是频繁项集）。

## 算法流程：

如何从大量的数据中找出不同项的关联规则呢？首先看一下Apriori实现过程：

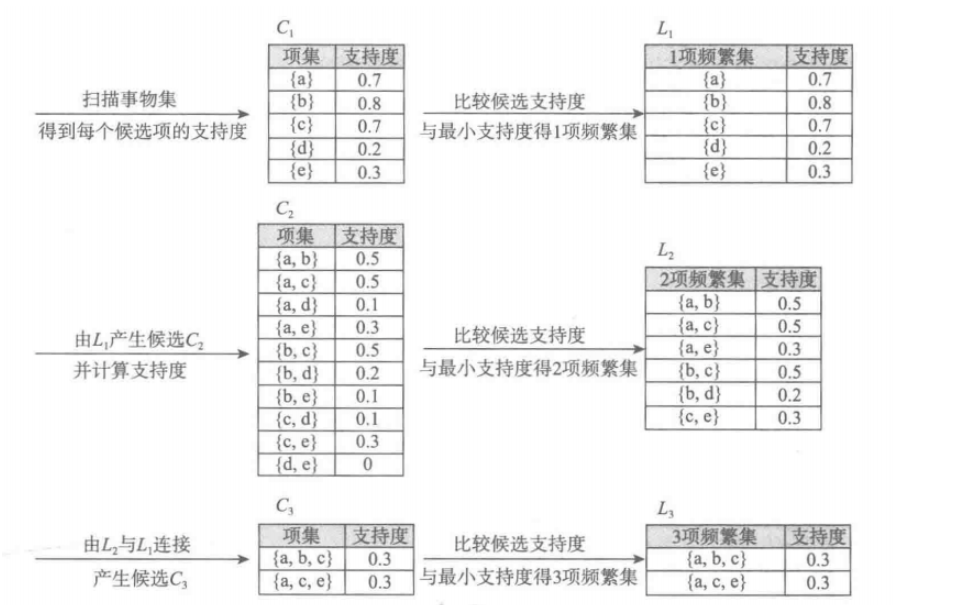
Apriori算法的主要思想是找出存在于事务数据集中最大的频繁项集，再利用得到的最大频繁项集与预先设定的最小置信度阈值生成强关联规则。



**Fig2：数据集**

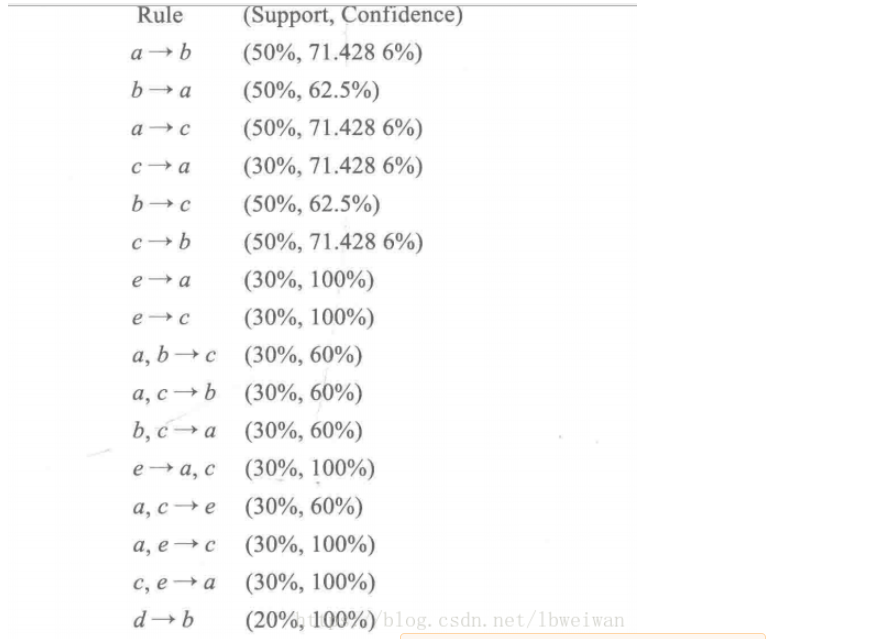
算法过程如下：

1. 首先计算出所有的频繁项集，这里最小支持度为0.2。



**Fig3：计算频繁项集**

得出L1、L2、L3的各个项集均为频繁项集，再进行计算每个频繁项集的置信度，其中L1不必计算。计算结果如下：



**Fig4：结果**