行业信息汇总模块详细设计

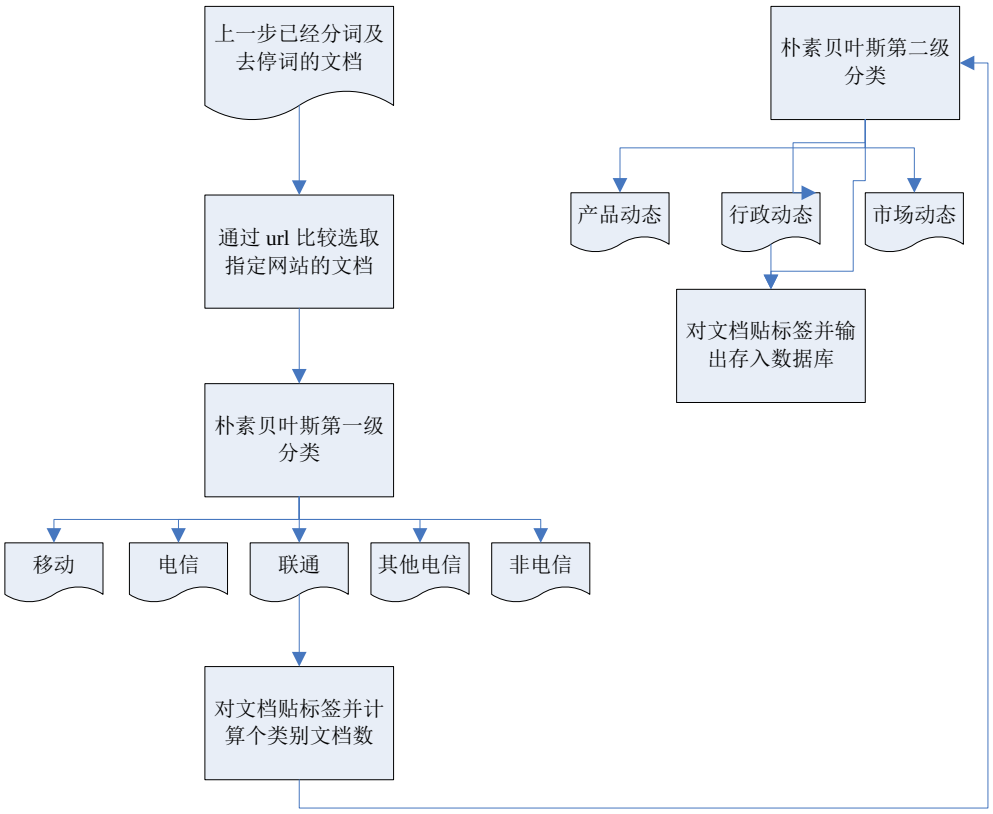
1. 提取指定网站的文档
2. 进行第一级分类，分为中国移动，中国电信，中国联通，其他电信，非电信
3. 进行第二级分类，分为产品动态，行政动态，市场动态

   产品动态：包括各个运营上在资费，套餐方面的新闻

   行政动态：包括各个运营上在人事，政策方面的新闻

   市场动态：包括各个运营上在网路建设，投资，合作伙伴，盈利状况方面的新闻

   总体流程图：

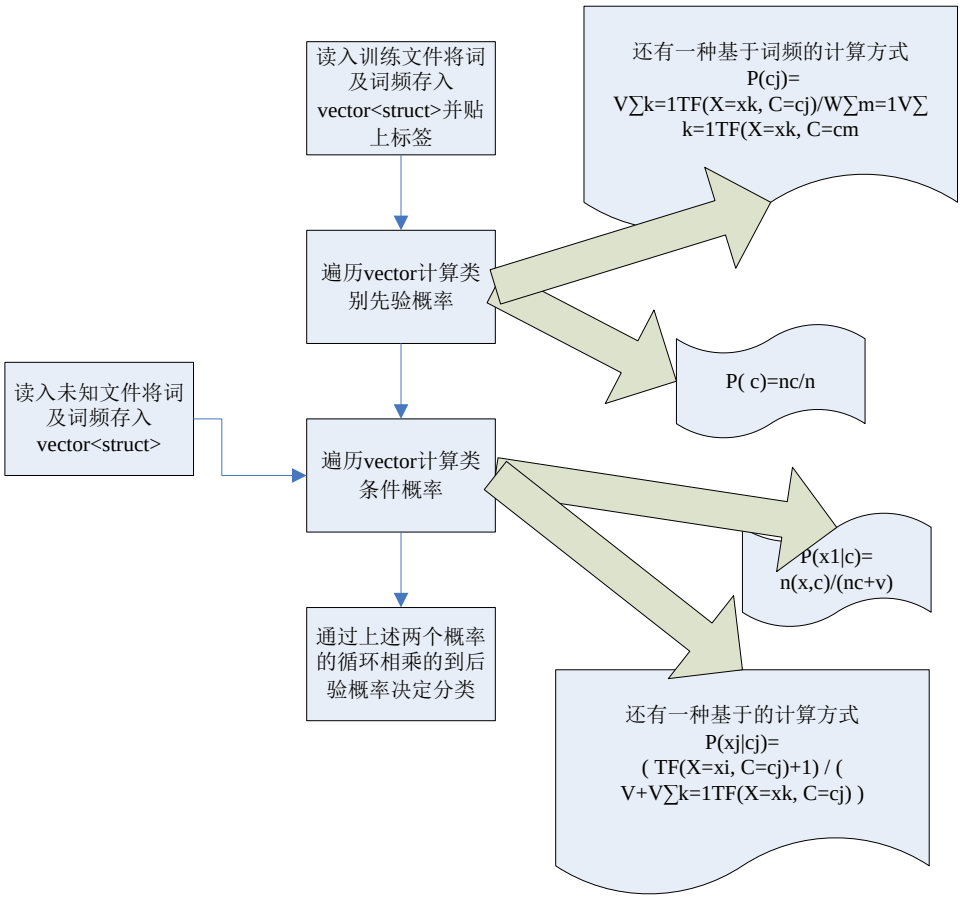


第一部分 提取指定网站的内容：

通过站点 url 和文档的 url 前面部分比较，得到指定的内容

这个可以把 url 存取成 string,通过 string 的函数就可以做得

第二部分 朴素贝叶斯分类：



这部分需要处理过得到的文档，需要进行分词及去停词操作。

首先，将预处理的文档读入写入，并计算词语的频度 TF 存成一个

Struct

{

词语，

频度

}

见整个文档存入 vector<struct>

并最后添加一个标签 struct

计算先验概率和类条件概率

计算后验概率=先验概率*类条件概率的循环

比较后验概率决定分类，存入数据库。