# 基于Aprori算法的医疗诊断规则的挖掘和分析

### 成员

组号：23组

* 严西敏 2120151051
* 欧阳童 2120151025

### Github 地址

https://github.com/yAnXImIN/MedicalApriori

### 处理步骤

1. 用数据爬虫从网站上面抓取网页。
2. 文本预处理，对文本数据进行去重和过滤掉那些没有实际意义的词。
3. 分词，使用专用的医疗学词库，对文本进行分词
4. 抽取 症状 和 疾病 信息
5. 使用Aprori算法，计算症状和疾病之间的相关性
6. 根据第5步的规则，输出可信度高的规则

其流程图如下所示：



**图 程序流程图**

### 项目简介

“医疗诊断专家系统”是指用于医疗诊断的专家系统，它把医学专家的疾病诊断与治疗经验，通过一定的计算方法编成程序，存入计算机内，然后应用计算机模仿医学专家，协助临床医生进行疾病的诊断与治疗，当病人就诊时，把病人的症状、体征及各种检查结果输入计算机，计算机按已存入的“经验”对疾病进行辅助诊断，并提出治疗建议。

我们知道人类专家之所以成为专家, 在于具备了在实践和书本中自动的提炼和积累知识的能力, 一个真正能够模拟人类专家的系统也应该具备这种不断获取知识的能力, 它或者提供一种使知识工程师和领域专家能够不断地给系统“ 传授”知识, 使知识库越来越丰富完善; 或者应能自动的或交互式地从原始资料、数据中提炼知识, 也就是自动机器学习。

应用于医学领域的专家系统知识库如果具备自我完善不断更新的能力，这对于诊断的准确信将会有很大的提升，知识的完善程度以及深度都会有改进，那么仅仅依靠领域专家不断给系统“传授”知识是不够的，需要系统能够从互联网广袤的资源中去学习发展。可以利用爬虫程序，定期的从权威医疗网站爬取数据，也可以利用现有的电子病历数据来作为原始数据输入，不断的更新知识库，让系统更为智能可靠。

下面，我们按照步骤来阐述我们的数据挖掘系统。

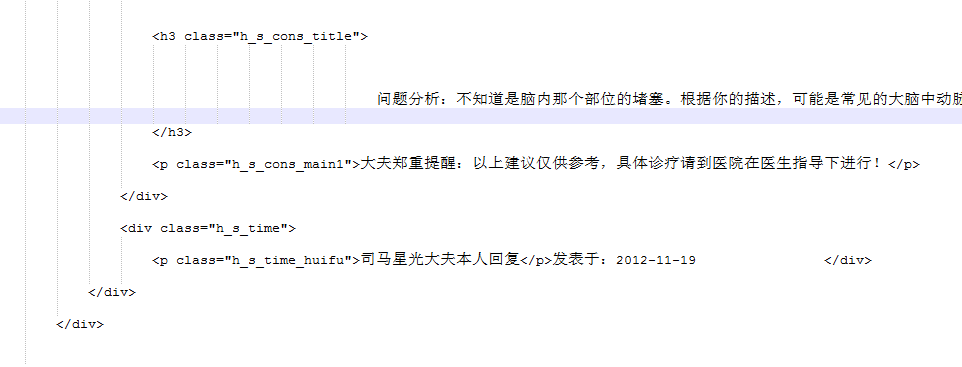
### 爬虫程序

首先是数据的获取，主要使用爬虫爬取：

利用爬虫工具，从较为权威、准确程度较高的医疗网站、医疗论坛等爬取内容，具体步骤如下：

1. 打开网页，分析网页原码中各个标签的属性，利用正则表达式进行匹配查找，找到分别包含患者咨询的问题与医生的回答情况的标签。
2. 分别将这两部分内容提取出来，保存到两个文档中。
3. 使用广度优先搜索的方式重复上述步骤，其中根据网址的不同过滤掉不合理的网页，重复多次直到数据量足够大。

最后获得的数据结果如下图所示：



**图 使用爬虫程序获得文本资料**

### 数据预处理

然后对上述数据进行预处理。在大数据获得的同时系统已经按照正则表达式对有效数据进行了提取，但是这些数据仍不能被很好的保存和分析。由于用户通过网络进行自主询问病情，大部分用户并不具备专业的医疗知识，所呈现在网页上的文本口语化严重，而且包含了大部分与所需数据无关的连接用语等，还有更多的是对症状描述不清，甚至使用俗语方言的情况，所以需要利用自然语言处理的相关知识对初始数据进行预处理的操作，得到专业的、系统可识别的数据并保存。步骤如下：

建立症状词典，将每个症状词汇按照[名称 属性 词频]的格式存储为一条记录，然后保存在后缀为dic的文件中，作为症状词典。

建立疾病词典，将每个疾病词汇按照[名称 属性 词频]的格式存储为一条记录，然后保存在后缀为dic的文件中，作为疾病词典。

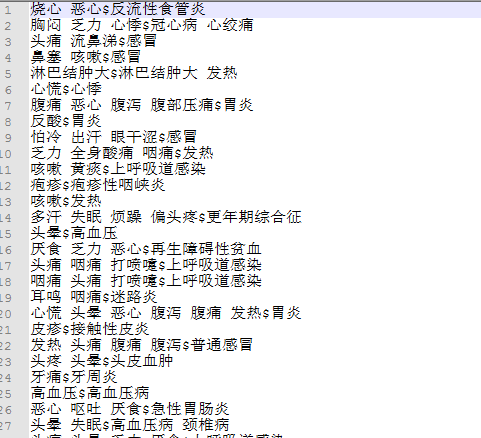
建立偏义词典，将每个症状、疾病词汇的近义词、俗语等按照[名称 对应词 词频]的格式存储为一条记录，然后保存在后缀为dic的文件中，作为偏义词典。

采用了开源Ansj中文分词方法，利用症状词典对之前获得的症状相关内容进行分词处理，同时利用疾病词典对之前获得的诊断结果相关内容进行分词处理，保证症状和结果一一对应的顺序不变。

为在以上两个词典中出现的词语可以再利用偏义词典进行处理，提高准确率，匹配之后提取对应的专业术语进行使用。

实现了对文本的分词处理之后，每个得到的词语看成一个数据点，每次进行分析之后所得到的数据点分成两部分，即症状和结果，总体上也就组成了一个二维的向量。进进行大量的数据处理之后就可以得到足够数量的二维向量向下传递，供接下来进行分析。

最后获得的结果如下图所示。



**图 使用分词程序和词典过滤处理之后的结果**

### 使用Apriori算法进行数据挖掘

在本方法中，数据库中存放的信息为每次医疗问诊过程中的病情描述和疾病诊断的已经经过处理的，已经被分词，过滤无关信息的数据条目。我们可以利用关联规则集挖掘的一些算法，来对这些数据进行处理。从而找出我们所需要的疾病症状到疾病诊断的关联规则集。最后可以将这些规则集放入专家系统中，进行行一步的分析利用。

关联规则就是在给定训练项集上频繁出现的项集与项集之间的一种紧密的联系。其中“频繁”是由人为设定的一个阈值即支持度（support）来衡量，“紧密”也是由人为设定的一个关联阈值即置信度（confidence）来衡量的。这两种度量标准是频繁项集挖掘中两个至关重要的因素，也是挖掘算法的关键所在。

其具体步骤如下所示

1. 扫描数据集（上文得出的二维向量），找出症状频繁1项集。
2. 从第二次循环开始，利用频繁k-1项集和频繁1项集的笛卡尔积结果，找出症状频繁k项集。重复此步骤，直到找不出症状频繁k项集为止。
3. 扫描数据集，找出疾病频繁1项集。
4. 将疾病的频繁1项集与症状频繁1至k项集做笛卡尔积，根据该笛卡尔积结果，找出疾病与症状集的频繁项集。

在4上生成的集合中，使用置性度阈值，将高置性度的疾病与症状集的频繁项集。该频繁项集即为需要的症状到疾病的一条规则，输出该规则。

通过以上的步骤，我们就可以找出高置性度和高支持度的症状到疾病的规则集。该规则集即是我们需要从大数据中挖掘出来的关联规则集。通过该规则集的产生，我们就可以使用该规则集来进行简单的疾病诊断等工作。

下面来简述各步骤使用的算法。

1. 步骤1中，依次遍历从数据库中读取到的症状条目，计算每个症状出现的次数，如果出现次数大于我们设定的阈值，则加入到症状频繁1项集中。
2. 步骤2中，将频繁k-1项集和频繁1项集做笛卡尔积操作，去掉无用的条目，计算各条目出现的次数，如果出现次数大于我们设定的阈值，则加入到症状频繁K项集中。
3. 步骤3中，依次遍历从数据库中读取到的疾病条目，计算每个疾病出现的次数，如果出现次数大于我们设定的阈值，则加入到疾病频繁1项集中。
4. 步骤4中，将疾病的频繁1项集与症状频繁1到k项集做笛卡尔积，根据该笛卡尔积结果，找出疾病与症状集的频繁项集。
5. 步骤5中，在步骤4中得到的频繁疾病与症状集之后，计算每个项目中的置信度，即在该症状集合中，该疾病出现的条件概率。

根据以上描述的基于Aprori算法的关联规则集挖掘方法，在以上各步骤中，利用频繁k-1项集找出症状频繁k项集，由于要判断集合是否相等，会极大地增加计算时间，提出了集合对比运算优化策略，如下所示。

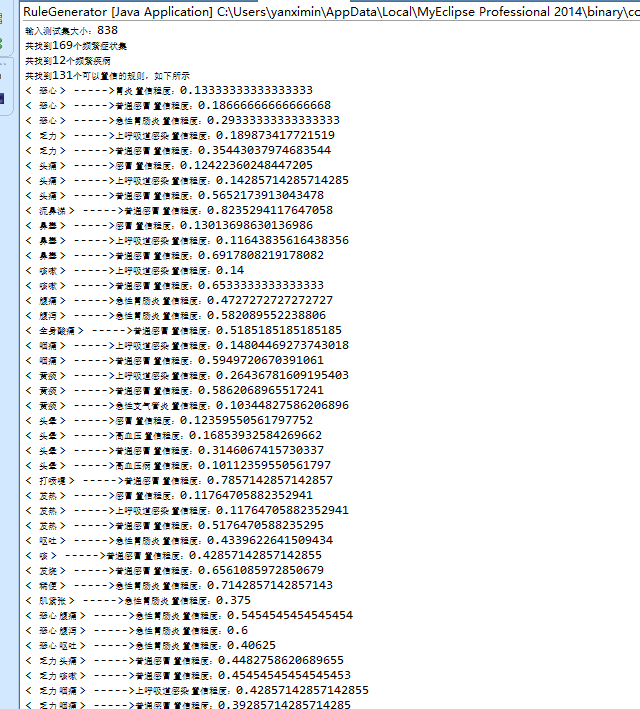
对于集合A和集合B，如果Hash(A)=Hash(B)，则认为集合A和集合B是相等的。

对于集合A和集合B，如果Hash(A)=Hash(C)，其中C为B的真子集，则认为A为B的真子集。

这里，对于集合的Hash函数定义如下：

HashStr是用来计算一个字符串的Hash值，定义如下：

上面公式中的lengthOf(A)即为字符串A的长度。Wi为字符串中的第i个字符。



**图 最后获得的规则集**

### 工作实例

|  |  |  |
| --- | --- | --- |
| 爬虫数据 | 问题：嗓子有痰但咳不出，嗓子痒一咳就憋的不能出气。 | 分词结果：有痰、咳 |
| 回答：您这种情况多久了？有没有诱因。根据你的情况是考虑上呼吸道感染的情况 | 分词结果：上呼吸道感染 |
| 组成向量：[有痰 咳 & 上呼吸道感染] | 说明：&区分症状和结果 |
| 病历数据 | 病人主诉：鼻流黄涕，咳嗽，咯吐黄痰，一周，现气喘，可闻及痰鸣音。 | 分词结果：咳嗽、黄痰 |
| 就诊意见：上呼吸道感染 | 分词结果：上呼吸道感染 |
| 组成向量：[咳嗽 黄痰 & 上呼吸道感染] |  |
| 规则挖掘向量示意： | 全身酸痛 鼻塞 咽痛 咳嗽 发烧$普通感冒  咽痛 咳嗽 黄痰 全身酸痛 流鼻涕$普通感冒  头痛 咽痛 淋巴结肿大$上呼吸道感染  头痛 乏力 咳嗽 发烧$普通感冒  头痛 全身酸痛 咳嗽 出汗$普通感冒  头晕 咽痛 咳嗽 发烧$普通感冒  咽痛 咳嗽$普通感冒  发热 全身酸痛$普通感冒  头痛 发热 咳嗽$普通感冒  咳嗽 黄痰 咽痛$普通感冒  鼻塞 头痛 咽痛 咳嗽$普通感冒  抽搐 高热 惊厥$高热  咳嗽 黄痰 咳 发烧$普通感冒  咳嗽$急性咽喉炎 |  |
| 规则挖掘结果： | < 腹泻 发烧 稀便 > ----->急性胃肠炎 置信程度：0.9166666666666666  < 咽痛 黄痰 发烧 > ----->普通感冒 置信程度：0.7333333333333333  < 咽痛 打喷嚏 发烧 > ----->普通感冒 置信程度：0.9090909090909091  < 流鼻涕 咳嗽 咽痛 发烧 > ----->普通感冒 置信程度：0.8571428571428571  < 鼻塞 咳嗽 咽痛 发烧 > ----->普通感冒 置信程度：1.0  < 咳嗽 咽痛 黄痰 发烧 > ----->普通感冒 置信程度：0.7333333333333333 |  |
| 系统使用： | 输入：您好，请问我最近有点咳嗽，还流鼻涕，请问我怎么了？ | 分词结果：咳嗽、流鼻涕 |
| 系统回答：可能为普通感冒 |  |

### 结果分析

我们对测试集大小为838的数据集进行挖掘，设置最小支持度为0.01，最小置性度为0.5，挖掘出来的规则如下所示：

输入测试集大小：838

共找到169个频繁症状集

共找到12个频繁疾病

共找到88个可以置信的规则，如下所示

< 头痛 > ----->普通感冒 置信程度：0.5652173913043478

< 流鼻涕 > ----->普通感冒 置信程度：0.8235294117647058

< 鼻塞 > ----->普通感冒 置信程度：0.6917808219178082

< 咳嗽 > ----->普通感冒 置信程度：0.6533333333333333

< 腹泻 > ----->急性胃肠炎 置信程度：0.582089552238806

< 全身酸痛 > ----->普通感冒 置信程度：0.5185185185185185

< 咽痛 > ----->普通感冒 置信程度：0.5949720670391061

< 黄痰 > ----->普通感冒 置信程度：0.5862068965517241

< 打喷嚏 > ----->普通感冒 置信程度：0.7857142857142857

< 发热 > ----->普通感冒 置信程度：0.5176470588235295

< 发烧 > ----->普通感冒 置信程度：0.6561085972850679

< 稀便 > ----->急性胃肠炎 置信程度：0.7142857142857143

< 恶心 腹痛 > ----->急性胃肠炎 置信程度：0.5454545454545454

< 恶心 腹泻 > ----->急性胃肠炎 置信程度：0.6

< 头痛 流鼻涕 > ----->普通感冒 置信程度：0.75

< 头痛 鼻塞 > ----->普通感冒 置信程度：0.7058823529411765

< 头痛 咳嗽 > ----->普通感冒 置信程度：0.7333333333333333

< 头痛 咽痛 > ----->普通感冒 置信程度：0.5633802816901409

< 头痛 发热 > ----->普通感冒 置信程度：0.68

< 头痛 发烧 > ----->普通感冒 置信程度：0.7647058823529411

< 流鼻涕 鼻塞 > ----->普通感冒 置信程度：0.92

< 流鼻涕 咳嗽 > ----->普通感冒 置信程度：0.8148148148148148

< 流鼻涕 咽痛 > ----->普通感冒 置信程度：0.8055555555555556

< 流鼻涕 黄痰 > ----->普通感冒 置信程度：0.7142857142857143

< 流鼻涕 打喷嚏 > ----->普通感冒 置信程度：0.8421052631578947

< 流鼻涕 发烧 > ----->普通感冒 置信程度：0.896551724137931

< 鼻塞 咳嗽 > ----->普通感冒 置信程度：0.6956521739130435

< 鼻塞 咽痛 > ----->普通感冒 置信程度：0.6666666666666666

< 鼻塞 打喷嚏 > ----->普通感冒 置信程度：0.9

< 鼻塞 发烧 > ----->普通感冒 置信程度：0.9032258064516129

< 咳嗽 咽痛 > ----->普通感冒 置信程度：0.6666666666666666

< 咳嗽 黄痰 > ----->普通感冒 置信程度：0.5813953488372093

< 咳嗽 头晕 > ----->普通感冒 置信程度：0.56

< 咳嗽 打喷嚏 > ----->普通感冒 置信程度：0.8064516129032258

< 咳嗽 发热 > ----->普通感冒 置信程度：0.5909090909090909

< 咳嗽 发烧 > ----->普通感冒 置信程度：0.75

< 腹痛 腹泻 > ----->急性胃肠炎 置信程度：0.5714285714285714

< 腹痛 呕吐 > ----->急性胃肠炎 置信程度：0.6666666666666666

< 腹痛 稀便 > ----->急性胃肠炎 置信程度：0.6111111111111112

< 腹泻 呕吐 > ----->急性胃肠炎 置信程度：0.68

< 腹泻 发烧 > ----->急性胃肠炎 置信程度：0.631578947368421

< 腹泻 稀便 > ----->急性胃肠炎 置信程度：0.7307692307692307

< 腹泻 肌紧张 > ----->急性胃肠炎 置信程度：0.5882352941176471

< 咽痛 黄痰 > ----->普通感冒 置信程度：0.6222222222222222

< 咽痛 头晕 > ----->普通感冒 置信程度：0.5833333333333334

< 咽痛 打喷嚏 > ----->普通感冒 置信程度：0.8064516129032258

< 咽痛 发热 > ----->普通感冒 置信程度：0.5277777777777778

< 咽痛 发烧 > ----->普通感冒 置信程度：0.71

< 黄痰 发烧 > ----->普通感冒 置信程度：0.6875

< 头晕 发烧 > ----->普通感冒 置信程度：0.7142857142857143

< 打喷嚏 发烧 > ----->普通感冒 置信程度：0.9047619047619048

< 呕吐 稀便 > ----->急性胃肠炎 置信程度：0.7692307692307693

< 发烧 稀便 > ----->急性胃肠炎 置信程度：0.9166666666666666

< 恶心 腹痛 腹泻 > ----->急性胃肠炎 置信程度：0.6666666666666666

< 头痛 鼻塞 咳嗽 > ----->普通感冒 置信程度：0.7727272727272727

< 头痛 鼻塞 咽痛 > ----->普通感冒 置信程度：0.6666666666666666

< 头痛 咳嗽 咽痛 > ----->普通感冒 置信程度：0.6842105263157895

< 头痛 咳嗽 发热 > ----->普通感冒 置信程度：0.8181818181818182

< 头痛 咳嗽 发烧 > ----->普通感冒 置信程度：0.8636363636363636

< 头痛 咽痛 发烧 > ----->普通感冒 置信程度：0.8

< 流鼻涕 鼻塞 咳嗽 > ----->普通感冒 置信程度：0.8947368421052632

< 流鼻涕 咳嗽 咽痛 > ----->普通感冒 置信程度：0.7878787878787878

< 流鼻涕 咳嗽 黄痰 > ----->普通感冒 置信程度：0.7142857142857143

< 流鼻涕 咳嗽 打喷嚏 > ----->普通感冒 置信程度：0.8125

< 流鼻涕 咳嗽 发烧 > ----->普通感冒 置信程度：0.85

< 流鼻涕 咽痛 打喷嚏 > ----->普通感冒 置信程度：0.9

< 流鼻涕 咽痛 发烧 > ----->普通感冒 置信程度：0.875

< 流鼻涕 打喷嚏 发烧 > ----->普通感冒 置信程度：0.9

< 鼻塞 咳嗽 咽痛 > ----->普通感冒 置信程度：0.6904761904761905

< 鼻塞 咳嗽 发烧 > ----->普通感冒 置信程度：0.9130434782608695

< 鼻塞 咽痛 发烧 > ----->普通感冒 置信程度：0.9333333333333333

< 咳嗽 咽痛 黄痰 > ----->普通感冒 置信程度：0.6222222222222222

< 咳嗽 咽痛 打喷嚏 > ----->普通感冒 置信程度：0.85

< 咳嗽 咽痛 发烧 > ----->普通感冒 置信程度：0.7349397590361446

< 咳嗽 黄痰 发烧 > ----->普通感冒 置信程度：0.6875

< 咳嗽 打喷嚏 发烧 > ----->普通感冒 置信程度：0.8571428571428571

< 腹痛 腹泻 呕吐 > ----->急性胃肠炎 置信程度：0.7333333333333333

< 腹痛 腹泻 稀便 > ----->急性胃肠炎 置信程度：0.6470588235294118

< 腹泻 呕吐 稀便 > ----->急性胃肠炎 置信程度：0.8181818181818182

< 腹泻 发烧 稀便 > ----->急性胃肠炎 置信程度：0.9166666666666666

< 咽痛 黄痰 发烧 > ----->普通感冒 置信程度：0.7333333333333333

< 咽痛 打喷嚏 发烧 > ----->普通感冒 置信程度：0.9090909090909091

< 头痛 鼻塞 咳嗽 咽痛 > ----->普通感冒 置信程度：0.75

< 头痛 咳嗽 咽痛 发烧 > ----->普通感冒 置信程度：0.7692307692307693

< 流鼻涕 咳嗽 咽痛 打喷嚏 > ----->普通感冒 置信程度：0.9

< 流鼻涕 咳嗽 咽痛 发烧 > ----->普通感冒 置信程度：0.8571428571428571

< 鼻塞 咳嗽 咽痛 发烧 > ----->普通感冒 置信程度：1.0

< 咳嗽 咽痛 黄痰 发烧 > ----->普通感冒 置信程度：0.7333333333333333

程序所用时间：413毫秒

### 结论

可以看到，从上述网站上获得的数据的可靠性还是比较高的。我们的算法可以很好的挖掘出一些可靠的令人置信的规则。