关联分析系统

## 1简介

### 1.1目的

本文档主要从使用说明以及架构方面对关联分析系统进行综合概述，其中主要包括系统的使用说明，总体架构，详细设计等。

### 1.2范围

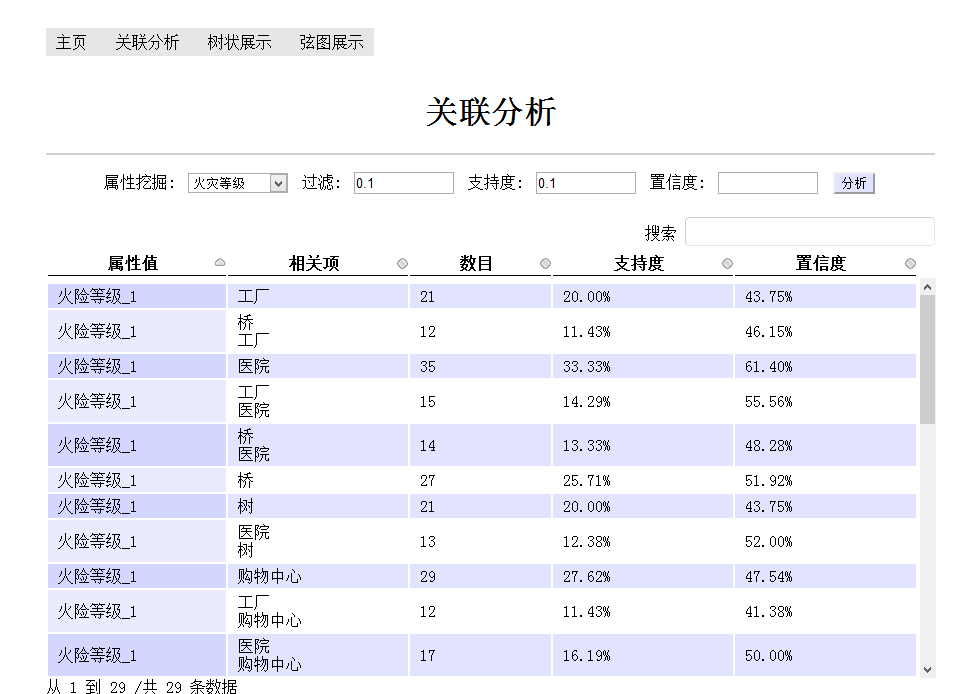
本文档从架构以及详细设计方队对关联分析系统做出描述，方便后续的开发与维护工作。

## 2使用说明

### 2.1简介

从用户的角度看，该系统还是比较简单的，一共只有4个页面；即主页，关联分析，树状展示和弦图展示。其中主页没有任何实质性的功能，仅仅是一些简单的文字说明。下面具体介绍其他三个页面的功能以及使用方法。

### 2.2关联分析



该页面的主要功能就是对数据按照用户设定的参数进行挖掘并以列表的形式进行展示，该页面一共接收4个参数，除了属性挖掘意外，其他三个参数的默认值都是0（警告：如果使用默认值0的话，将导致挖掘过程非常缓慢，同时还将产生大量的无用数据）。这4个参数的具体含义如下：

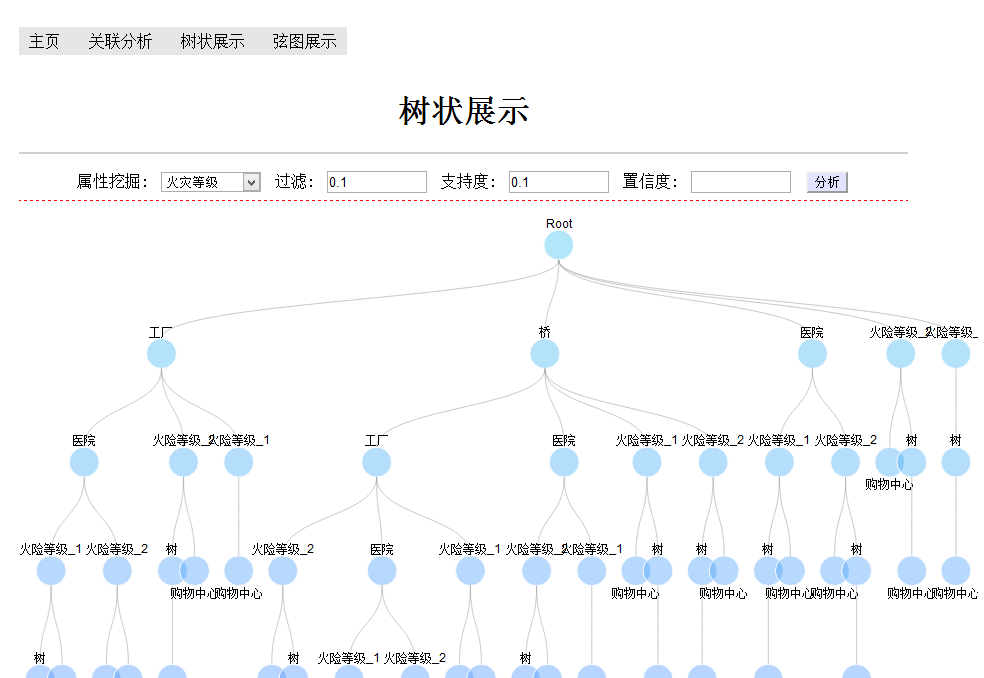
属性挖据：数据挖据的分类属性，比如设置为“火灾等级”，那么该模块就会按照“火灾等级”对所有的数据进行关联分析。

过滤：关联规则的最小差值，主要用来对挖掘的结果进行过滤；假设火灾一共有三个等级，即一，二，三；且有关联规则“工厂 饭店”，只有当该规则导致火灾等级一，二，三之间的概率差值均大于用户设定的值时，才认为这个规则是符合要求的，其取值范围为[0,1]。

支持度：同样是对挖掘的结果进行过滤；这是相对于分类属性中的某个具体值而言的，例如“火灾等级一”，那么某条规则对“火灾等级一”的支持度就是该规则在所有导致“火灾等级一”的规则中所占的比例，只有当这个比例不小于用户设定的值时，才认为其符合要求，其取值范围为[0,1]。

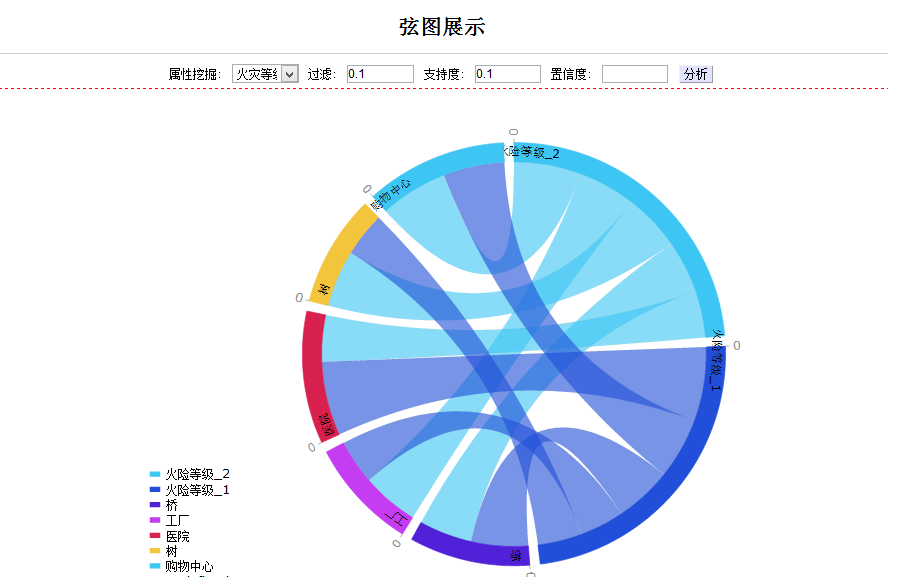
置信度：也是对挖掘的结果进行过滤；置信度是相对于分类属性的某个具体值而言的，例如“火灾等级一”的置信度就表示该规则导致的所有火灾中，“火灾等级一”所占的比例。 只有当某条规则的置信度不小于用户设定的值时，才认为其符合要求，其取值范围为[0,1]。

### 2.3树状展示



该页面用来展示数据FP-Tree算法的结果，从任意一个节点往上走到树根部位（不含树根），都是一条关联规则。

### 2.4弦图展示



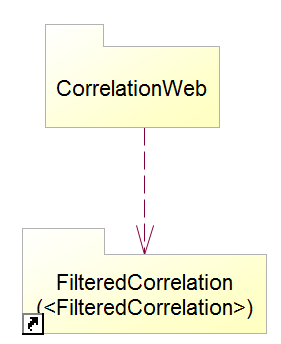
该页面的展示的是数据挖掘的最终结果，圆环上的每段弧线代表一个属性，不同属性之间的连线则表示它们之间存在关联关系。

## 3详细设计

### 3.1架构设计

为了降低各不同职能模块之间的耦合性，本系统将核心算法模块与用户界面显示模块进行了分离，用户界面模块通过接口调用核心算法模块，从而极大地降低了两个模块的耦合性，便与两个模块的并行开发以及维护。

### 3.2架构简图



#### 3.2.1 FilteredCorrelation

##### 3.2.1.1简介

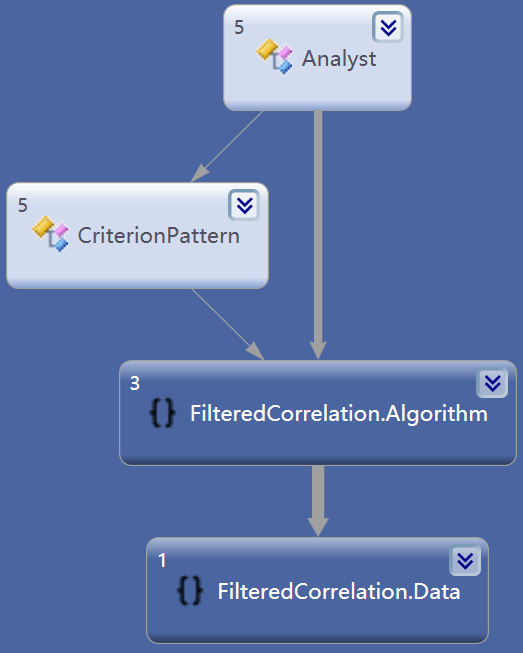
该模块是关联分析系统的核心模块，该模块主要实现了关联分析算法，同时还包括本系统与数据库的交互模块。

##### 3.2.1.2接口介绍

本系统的关联分析功能完全由该模块提供。在进行关联分析时，该模块一共接收4个参数，即

1. criterionAttr：对应用户界面的“属性挖掘”参数。
2. diffRatio：对应用户界面的“过滤”参数。
3. support：对应用户界面的“支持度”参数。
4. confidence：对应用户界面的“置信度”参数。

##### 3.2.1.3模块内部架构



###### 3.2.1.3.1 Analyst

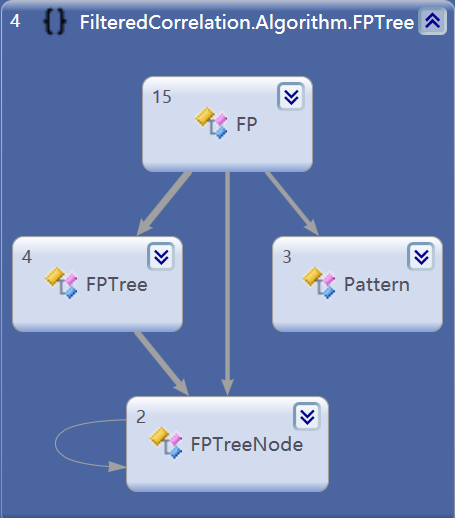
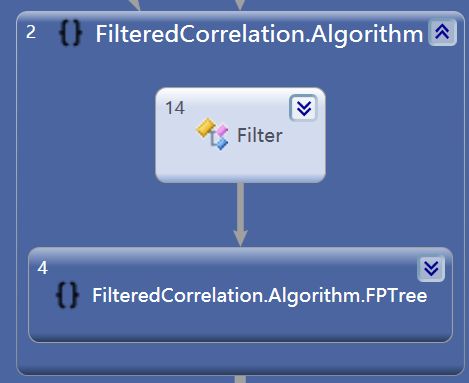
Analyst是该模块总的控制类，当外部模块和该模块进行交互时，都是通过Analyst来完成。上面提到的四个参数也是通过Analyst传给FilteredCorrelation模块。通过一系列的运算之后，返回给外部调用者一个CriterionPattern的列表。

###### 3.2.1.3.2 CriterionPattern

CriterionPattern表示一条关联规则。

###### 3.2.1.3.3 FilteredCorrelation.Algorithm

该模块主要是对数据集进行关联规则挖掘，并按照设定的过滤参数对结果进行过滤。下图展示该模块的架构。

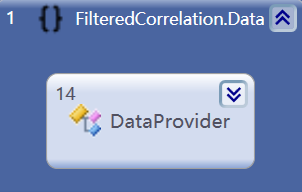


Filter类的作用就是按照设定的过滤参数对数据集以及挖掘出来的结果进行过滤，只保留符合要求的规则。这么做的好处是一方面可以去除无用的数据，减少数据挖据时的数据集规模，从而加快数据挖掘速度；另一方面，也可以去掉一些特征并不是那么明显的规则，从而减少干扰项。

FilteredCorrelation.Algorithm.FPTree，这个包是整个系统的核心所在，它的作用就是运用FP-Tree算法，对数据集进行关联规则挖掘。右图就是这个包里面的具体架构。这个算法主要运用到的数据结构是一个树状结构，其中的每个节点由FPTreeNode表示。Pattern则表示一条关联规则，它和CriterionPattern的区别在于CriterionPattern包含了更多信息，以便外部调用者能够更好地理解每条规则的具体含义。

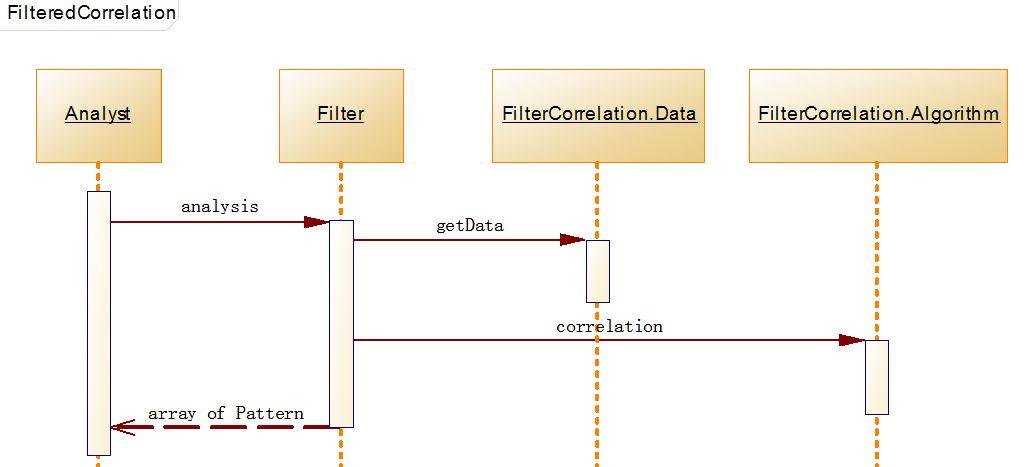
###### 3.2.1.3.4 FilteredCorrelation.Data

该模块是关联分析系统的数据访问模块，主要为本系统提供数据访问服务。这个包内部的结构比较简单，只含有一个类即DataProvider。



###### 3.2.1.3.5执行流程

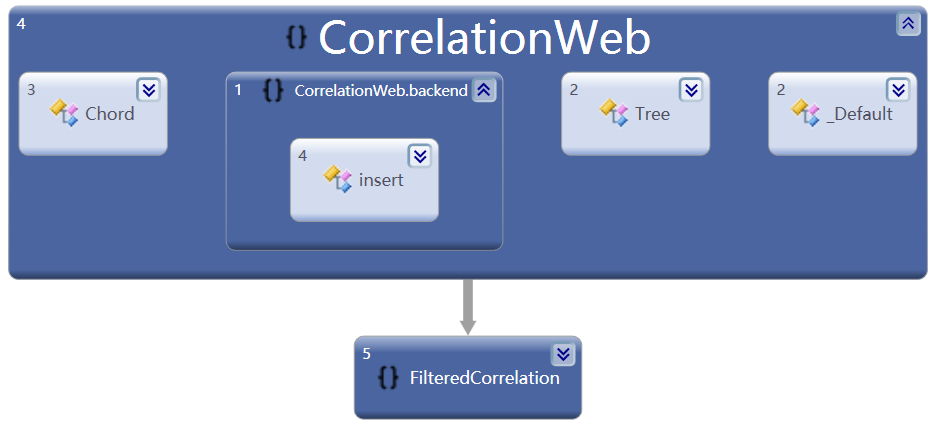
下图简要展示了一个关联分析挖掘的完整过程。



首先Analyst接收到外部调用者的传递参数之后，会调用Filter，并将过滤参数传递给Filter类；Filter类通过FilteredCorrelation.Data这个包获取到数据集之后，首先会对数据集进行一次过滤，然后将这些数据传递给FilteredCorrelation.Algorithm进行挖掘；待挖据完成之后，Filter将会对结果再进行一次过滤，最后将结果以Pattern数组的形式返回给Analyst。Analyst最后将Pattern包装成CriterionPattern的形式返回给外部调用者。

#### 3.2.2 CorrelationWeb

CorrelationWeb是关联分析系统的web显示模块。它的作用就是将用户的输入转化为FilteredCorrelation模块可接收的参数，并传递给FilteredCorrelation，最后将其返回的结果显示在网页上。其架构如下：



Chord表示显示为弦图的页面所对应的处理类。

Tree则表示显示为树状图的页面所对应的处理类。

\_Default表示显示为关联列表页面所对应的处理类。

CorrelationWeb.backend则是用来给数据库插入随机数据。