四川大学

学士学位论文

基于频繁模式挖掘的网络异常事件检测系统

学生姓名：李盛乐

学科专业：计算机科学与技术

指导导师：周颖杰

完成时间：哈哈哈

SiChuan University

A dissertation for bachelor's degree

Network abnormal event detection system based on frequent pattern mining

Student’s Name:Shengle Li

Speciality:Computer science and technology

Supervisor:Yingjie Zhou

Finished time:emmmm…

摘要：随着互联网时代的到来，特别是移动互联网的高速发展，web服务和微服务已经深入到人们生活的衣食住行，人们使用互联网进行搜索、购物、支付、娱乐等等。因此，保障web服务的稳定运行变得十分重要。

Web服务的稳定运行主要靠运维人员来保障，运维人员通过监控各个关键性能指标（KPI）来判断web服务是否发生异常，KPI 异常的发生往往意味着相关的应用发生了异常。面对大量的KPI数据，亟需通过高效的网络异常事件检测系统帮助运维人员检测网络异常事件。本论文就围绕着如何通过频繁模式挖掘的算法建立一个网络异常事件检测系统，辅助运维人员对网络异常的检测。其中包括对频繁模式挖掘算法的研究和对网络异常事件检测系统的设计研究，其主要研究内容和贡献包括：

1. **对频繁模式挖掘算法的研究和使用**

本论文使用的频繁模式挖掘算法为经典的Apriori算法，Apriori算法的核心思想是通过候选集生成和情节的向下检测两个阶段来挖掘产生频繁项集，在算法执行的过程中伴随着剪枝的操作，最后通过最小支持度得到频繁项集。

1. **网络异常事件检测系统的设计和实现**

本毕业设计最终设计和实现一个网络异常事件检测系统，该系统的主要功能包括：（1）选择训练集和测试集数据；（2）绘出指定KPI的曲线图和异常点，以及异常的占比；（3）对训练集进行训练后，可选择不同类型的KPI进行检测；（4）计算并展示检测的精确率和召回率。

关键词：频繁模式；Apriori；数据挖掘；网络异常；关键性能指标（KPI）

Abstract: With the arrival of the Internet age, especially the rapid development of mobile Internet, web services and micro-services have gone deep into people's lives. People use the Internet for search, shopping, payment, entertainment and so on. Therefore, it is very important to ensure the stable operation of web services.

The stable operation of web services is mainly guaranteed by the operation and maintenance personnel. The operation and maintenance personnel determine whether the web service is abnormal by monitoring each key performance indicator (KPI). The occurrence of an abnormal KPI often means that the related application has an exception. Faced with a large number of KPI data, it is urgent to use an efficient network anomaly event detection system to help operation and maintenance personnel detect network anomalies. This dissertation focuses on how to establish a network anomaly event detection system through the algorithm of frequent pattern mining to assist network operators in detecting network anomalies. Including the study of frequent pattern mining algorithms and the design of network anomaly event detection system, its main research contents and contributions include:

1. Research and use of frequent pattern mining algorithms

The frequent pattern mining algorithm used in this dissertation is the classic Apriori algorithm. The core idea of Apriori algorithm is to mine frequent item sets through the generation of candidate set generation and the downward detection of the plot. The algorithm is accompanied by pruning during the execution of the algorithm. The operation finally obtains frequent item sets with minimum support.

1. Design and implementation of network anomaly event detection system

The graduation design finally designs and implements a network anomaly event detection system. The main functions of the system include: (1) selecting training set and test set data; (2) plotting the specified KPI graphs and abnormal points, and accounting for abnormalities (3) After training the training set, different types of KPIs can be selected for detection; (4) Calculate and display the precision and recall rate of the test.

Key words: Frequent mode; Apriori; Data mining; Network abnormal; Key performance indicators

1. 绪论
   1. 研究背景及意义

随着计算机及网络技术的发展，网络服务逐渐发展到社会的各个领域，特别是互联网及移动互联网时代的到来，web服务更是与人们的衣食住行息息相关，人们使用互联网购物、搜索、支付和娱乐等等，都离不开web服务。新中国四大发明高铁、移动支付、共享单车和网购，其中就有三个和web服务紧密相关，如今的社会没有web服务是难以想象的。由此可见，web服务对于现在社会已经不可或缺，保障其稳定性、可用性和高效性十分的重要。与此同时，很多不法分子利用一些技术对网络和服务进行攻击，此类针对服务的流量攻击威胁问题十分严重，造成服务的瘫痪、网络的拥堵和信息泄露等很多严重的问题，蠕虫入侵、SYN洪水、特洛伊木马、DNS高速缓存污染等网络攻击频繁发生，对网络的安全性和web服务的稳定性造成了严重的危害。如何在高速网络环境下对网络异常行为进行及时感知和快速处理，对于保证网络有效运行和提高服务提供能力的稳健性具有非常重要的意义[1] 。

目前的网络异常检测手段主要是依靠运维人员实时监控各种各样的关键性能指标（KPI）来判断服务是否稳定，异常的KPI 通常意味着相应的应用服务出现了问题。KPI 数据可以分为服务KPI和机器KPI，如今应用服务每天需要处理海量的服务请求，由此产生的KPI数据十分庞大，面对大量的网络KPI数据，仅靠人工监控以无法满足对服务异常检测的需求。而如今大数据挖掘和分析技术以十分成熟，在交通、金融、天气等方面均有成功的应用，结合大数据技术设计网络异常检测算法有很大的研究意义。

网络异常检测是指通过学习运维人员指出的异常数据的特征，运用数据挖掘分析和深度学习的技术，不断优化检测的算法，找出最有的诊断方法，从而对测试数据进行检测。异常检测在现实中存在很多的问题和困难，导致其不能运用于实际场景中，主要的困难如下：

1.相对于正常数据，异常事件很少发生。在实际的web应用场景中，业务系统发生异常的时刻很好，因此可用于挖掘和分析的异常数据很少；

2.异常种类繁多。在实际的业务系统中，应用十分复杂，而且处于不断更新状态，所以运行期间产生的故障类型很多，导致异常的类型也多种多样；

3.KPI的多样性。KPI有多种表现的型式，有周期型、稳定型、不稳定型和持续波动型等；

4.正常和异常的界限不是十分的精确，某些KPI服务的正常和异常的界限难以划分，导致对异常的判断十分困难。

这些难点导致现有的检测算法准确率和召回率都不高，不仅没能真正的帮助运维人员检测异常，反而因为误报和漏报增加了运维人员的工作量。所以，研究和设计一个能够有效检测异常事件的系统，辅助运维人员在现实场景中检测web服务数据中的异常，对于减少运维人员的工作量和保证服务的稳定性十分重要。

* 1. 研究现状
     1. 异常检测算法现状
        1. 基于分类的检测算法

基于分类的异常检测算法可以通过标签的种类分为两类：多分类问题和单分类问题。基于分类的检测算法种类十分多，以下简单介绍一些常见的算法：

1. 基于神经网络

神经网络是指多个输入输出单元，这些单元之间相互连接，而且每个连接都关联一个权重。神经网络通常是由输入层、隐含层和输出层三层构成，其在单分类和多分类事件中均可使用，该异常检测算法的主要步骤分为两步：（1）通过训练集的数据训练一个神经网络来学习不同的分类，在学习的过程中不断修正网络连接节点的节点阈值和权重；（2）利用学习所得的神经网络对测试集的每个样例进行检测，判断其是否异常。

1. 基于贝叶斯网络

该算法是一种使用概率统计学知识进行分类的算法，其主要在多分类异常检测中使用。贝叶斯网络进行分类预测的主要原理是：对于一个单特征分类的数据集，利用朴素贝叶斯网络来预测所给的数据类的后验概率，对于测试数据，其标签类是后验概率最大的标签类。贝叶斯网络中的零概率问题，可以结合拉普拉斯平滑进行改良。还可以通过汇总每个属性的后验概率来推广到多特征数据集。

1. 基于支持向量机

该算法是通过建立一种二分类模型对数据进行检测，经常使用于单分类的问题中。支持向量机有一个基础模型，该模型是一个间隔最大的线性分类器，其定义在特征空间上。其实质上是一个非线性分类器，因为其包含核技巧。在测试数据中，如果测试集数据在学习生成的区域之外则为异常点。

1. 基于规则

该算法是需要通过测试集学习定义一个样例为异常点的规则。如果测试数据不包含在学习出来的任何一个规则中则被定义为异常点。通常的步骤分为两步：（1）通过训练集挖掘异常模式的频繁项集，且这些频繁项集在正常模式中不频繁出现；（2）得到频繁项集后，测试集中符合频繁项集的样例被判断为异常数据。

基于分类的检测算法种类很多，算法的适用场景和算法优势也各有不同，它们也有各自的不足，其优势和不足如下：

（1）此类算法不仅能解决单分类问题，也适用于多分类问题，在测试建立模型时算法较快；

（2）多分类问题需要精确的分类标签，直接将一个准确的标签赋给测试数据有时候不准确。

* + - 1. 基于最近邻的异常检测算法

该算法的原理是检验所有的样例，计算不同样例之间的距离大小，得到所有样例数据后，对比不同样例的距离，将远离大部分样例数据较远的点判断为异常。此类算法有以下几种：

1. 基于到K个近邻的距离

通过计算每个点到与其距离最近的K个点的距离，然后累加起来，异常点通常是距离大部分数据比较远的点，这就是K近邻方法。

1. 基于相对密度

以一个数据为中心，寻找能够包含K个近邻的最小超平面，求其密度，与其临近的点的密度相似则为正常点，否则为异常点。

基于最近邻的异常检测算法的优缺点：

（1）适应于多种数据类型，只需要定义适合的距离即可；（2）算法的复杂度较大，不容易得到适合的间距。

* + - 1. 基于聚类的异常检测技术

基于聚类的算法通过设置小聚类簇的大小阈值、聚类簇距离阈值，舍弃远离其他聚类簇的小聚类簇，与其他对象非强相关的对象为异常对象。常用的聚类方法有K-means、EM、层次聚类算法等。

基于聚类的异常检测的优缺点：

（1）通过数据挖掘聚类的方法，在样本充足的情况下准确度相对较高；

（2）训练过程中的计算量较大，原理复杂；

（3）想要得到合理的模型，需使用大量的测试数据进行学习训练；

（4）如果异常数据单独成簇，则难以发现异常。

* + - 1. 统计异常检测技术

该算法的思想是，在一系列数据构成的随机模型中，正常点存在于高频率区域的可能性较大，异常数据点存在于低频率的可能性较大。统计异常检测技术常用的方法有两种，包括参数化方法和非参数化方法。

其中参数化方法有：

1. 基于混合参数分布；
2. 基于回归模型；
3. 基于高斯模型。

非参数化方法有：

1. 基于密度估计函数；
2. 基于核函数；
3. 基于直方图。

该类算法也有其适用的场景，以及自己的不足和优势，其特点如下：

1. 数据分布的假设难以确保成立，直方图难以发现属性间的关系；
2. 如果估计分布阶段对异常足够鲁棒，可以用无监督方式；
3. 如果数据分布的假设成立，会有统计上的异常检测方案。
   * + 1. 信息理论异常检测技术

信息理论异常检测技术的基本思想是数据中的异常会引起数据集信息内容的不规则性，通过一个数据的存在与否是否影响数据的规则性进行判断数据是否属于异常。

该技术主要的优势和不足如下：

1. 其不必像神经网络等算法要建立模型，对分布不必做出假设；
2. 效果的好坏取决于评估方法，难以得到数据的异常程度。
   * + 1. 谱异常检测技术

谱异常检测技术的基本思想是将数据嵌入到一个较低维的空间中，在此空间中正常情况和异常情况显得不同。判断异常样例的方法是寻找此空间的子空间。

谱异常检测技术的优势和不足如下：

1. 对于高维度的数据样例，可以降低数据的维度，使其在此类数据中也能使用；
2. 该算法在时间上的复杂度比较高。
   * + 1. 上下文异常检测技术

此类算法的思想是识别样例数据的上下文，然后计算测试样例的异常得分。算法的步骤是：

1. 通过样例数据的上下文属性识别其上下文；
2. 通过已知点的异常检测算法计算每个测试样例的异常得分。

该算法的主要优势和不足有：

1. 相关技术能够生效的前提是，上下文相关的属性得到界说；
2. 得到实际异常的定义后，可利用这些定义来识别测试样例的异常。

上述中的一些算法在网络异常事件检测中已有一定的研究和应用，但由于网络异常的种类多种多样且变化十分迅速，面对不同类型的网络异常，其中的某种算法难以达到较高的精确率和召回率，所以需要研究更加高效准确的算法用于网络异常检测系统中。

* + 1. 频繁模式挖掘算法现状

频繁模式挖掘是指通过算法挖掘出频繁出现在数据集中的模式，对数据集的每个项，计算其支持度和置信度，将置信度和支持度大于阈值的项集定义为频繁项集。再通过频繁项集找出对象间的关联关系。频繁模式挖掘算法多种多样，目前该挖掘算法已在图像研究领域、文本分类、网络入侵检测技术、信息检索查询和web推荐系统等领域中有所应用。

在网络异常领域中频繁模式挖掘算法研究和应用还较少，没有很好的研究成果，频繁模式挖掘算法通过挖掘数据中的频繁模式，并且这些模式在正常标签中不频繁出现，通过这些频繁模式对测试集进行预测，理论上能达到较好的效果，所以研究出频繁模式挖掘算法在网络异常检测中的应用可行度很高，具有很好的应用前途。

* 1. 本文研究内容

基于上述的背景和研究意义，本文主要研究频繁模式挖掘算法的原理和实现过程，并通过这些技术设计一个网络异常事件监测系统，对网络KPI数据进行检测，辅助运维人员对应用服务进行监控。

具体的研究内容为基于频繁模式挖掘的网络异常事件检测系统，具体工作分为：

1. 数据的收集和预处理：挖掘项集的频繁模式，找出关联规则，需要对大量的数据进行训练，通过训练数据得出频繁项集。从竞赛官网下载的关于各种类型KPI的数据存在一定的缺失，需要对数据进行预处理补全缺失值。
2. Apriori算法的研究和实现：本研究使用经典的频繁模式挖掘算法Apriori算法，该算法在简单易学，实现上也不存在很大的困难，而且算法中的连接和剪枝能够保证较快的速度和较高的准确率。
3. Matlab软件的使用学习：MATLAB是一个方便、高效、功能强大的开发集成环境，其在数据可视化、矩阵计算、算法开发等方面有很好的变现，并且其具有界面系统开发的功能，使用该工具能够快速开发一个良好的检测系统。
4. 系统功能和界面的设计和实现：本设计最终需要完成一个网络异常事件检测系统，需要设计系统的功能、数据流和界面等，并实现其中的功能、处理数据的流向并将界面和后台实现连接起来，完成一个易于使用且简洁优美的系统。
   1. 论文结构安排

本文分为四个章节，每个章节注重介绍的内容均不同，本文的大纲以及内容简介如下：

第一章是绪论，主要叙述了网络异常事件检测的研究的背景和意义，分析目前国内外的研究现状，简述主要的算法原理和使用场景，以及每个算法特点、优势和不足，介绍本研究主要的研究和工作内容，最后描述论文的结构和大纲的安排。

第二章主要介绍经典频繁模式挖掘算法，以及详细介绍本研究使用的Apriori算法的原理和步骤，描述研究的方法和过程，阐述研究和实验的结果和结论。

第三章主要阐述设计网络异常事件检测系统的工作过程，其中包括对设计系统进行的需求分析、系统的功能设计、系统数据处理流程和系统界面的设计和后台实现，展示最终完成的系统。

第四章是对本次毕业设计和研究工作的总结和展望，叙述本次毕业设计的收获、遇到的问题、解决的方法，以及对研究内容未来的展望。

1. 频繁模式挖掘算法研究
   1. 频繁模式挖掘算法介绍
      1. 概念

（引）频繁模式主要包括频繁项集、频繁子序列或者子结构，而频繁项集是频繁模式中最重要的研究内容，是生成关联规则的重要一步，也是其他频繁模式的依据。

频繁模式挖掘算法有很多种类，以及很多与其他算法相结合所得的算法变种。其中三种经典的频繁模式挖掘算法分别是Eclat算法、Apriori算法和FP-Growth算法，下面将详细讲解这三种算法的原理和实现过程，并对它们的优势、不足和适用场景进行分析。

* + 1. Apriori算法

在众多数据挖掘算法中，有一个重要的研究方向就是关联规则挖掘的算法研究，而Apriori算法是最经典的关联规则挖掘算法之一。算法首先预置一个最小的支持度和最小的置信度，计算每个项集的支持度，经由这两个阈值遴选频繁项集，如果一个项集不小于最小支持度，称其为频繁项集；如果一个规则不小于最小置信度，称其为强规则。

Apriori算法将关联规则挖掘任务分为两个步骤：

1. **频繁项集的产生**：第一步通过连接和剪枝两个方法得到频繁项集，扫描数据库得到所有的1-项集，计算每个项集的支持度，删除小于最小支持度的项集，删除不符合规则的项集的步骤叫做剪枝步；然后1-项集自身连接得到2-项集，此步骤成为连接步；重复剪枝步和连接步，直到不再产生新的k-项集；
2. **关联规则的产生**：第一步得到频繁项集后，通过这些频繁项集可以找出关联规则。每个频繁项集可以产生一些规则，计算每个规则的置信度，将置信度不小于阈值的规则成为强规则。

Apriori算法的两个重要概念是：

1. 如若一个项集是频繁项集，则它的所有子集是频繁项集；
2. 如若一个项集不是频繁项集，则它的所有超级也不是频繁项集。
   * 1. FP-Growth算法

FP-Growth算法使用一种自定义的数据结构来优化算法的速度，这种数据结构成为FP-tree，它包括频繁项头表和项前缀树两个部分。

FP-Growth算法分为两个步骤：

1. 构建一棵FP树。FP树是一种特殊的前缀树，由频繁项头表和项前缀树构成，其按支持度降序排序，支持度越高的频繁项离根节点越近，从而使得更多的频繁项可以共享前缀。在构造FP树的时候，算法共2次读取遍历数据文件，第一次遍历的作用是计算发生的频度，第二次遍历忽略非高频度的对象。
2. 找出（1）中得到的tree中的频繁项集。（1）中得到tree后，就能利用这些tree找出频繁项集，方法是：（1）通过tree得到数据中的条件模式基；（2）通过（1）的结果，生成一棵条件tree；（3）反复进行（1）和（2）步骤，到树指囊括单个对象结束。
   * 1. Eclat算法

Eclat算法的思想是：两个K项的频繁项集经过或运算可以生成K+1项的频繁项集，而且将这两个K项频繁项集按照从小到大排序后，这两个K项频繁项集除了最后一项不一样外，之前的K-1项是一样的。

依据上述理论，Eclat算法可以扫描一次数据库，先获得一维的频繁项集，然后直接处理一维的频繁项集递归得到所有的频繁项集，无需在扫描数据库。实现该算法有两个阶段：

1. input阶段，主要功能是扫描一次数据库，数据存入onemap中，再导入到mymap中；
2. connect阶段的作用是利用上个步骤得到的I频繁项集，产生I+1频繁项集。
   * 1. 三种算法对比
3. 三种算法中，Apriori算法的效率最低，为了计算项集的支持度和置信度，其需要数次遍历样例文件，遍历的过程花费很多的性能和时间；
4. FP-Growth算法在长事物数据上表现很差，因为随着事物的增加，树的深度也增加，需要求解的子问题变得更多，因此效率会下降；
5. Eclat算法的效率最高，但是由于Eclat算法使用递归的方法求解，所以数据量很大时，会给系统带来很大的负担，不适用与=于数据量大的情况。
   1. Apriori算法

上一个小节中已经介绍了频繁模式挖掘算法的基本概念和其三种经典算法，基于本文的研究方向为使用Apriori的算法进行网络异常模式挖掘，本小节将详细介绍Apriori算法的相关概念和算法原理。

* + 1. 基本概念

学习该算法首先需要了解很多基本的知识，例如支持度等一些知识。具体知识如下：

1. 支持度：两个事件S和T，关于S->T，这个规则的支持度的意思就是事件S发生的时候，事件T也发生的几率，计算方法是S和T同时发生的次数除以总得样例的次数；
2. 置信度：事件S和T，置信度(S|T)的意思是事件S发生时，事件T也发生的几率，计算的方法是事件S和T同时发生的次数除以事件S发生的总的次数。
3. 项集：项集是指样例中对象与对象的组合，将含有S个对象的样例成为S-项集。
4. 关联规则：关联规则是映射某个元素与其他元素内部的联系关系，是数据挖掘的一个重要技术，其可以得到难以察觉的元素与元素之间的关系。
   * 1. 算法步骤

以下为其操作的过程，主要分为连接和剪枝，具体如下：

1. 扫描数据集，对每个候选项集计数C1，计算候选项集的支持度，删除支持度小于最支持度的项集得到频繁1-项集合L1；
2. 通过频繁K-项集生成频繁K+1-项集，方法是将所有频繁K-项集按照从小到大排序，取任意两个只有最后一个元素不同的频繁K-项集求取并集得到所有的频繁K+1-候选项集，计算所有频繁K+1-候选项集的支持度，删除不满足最小支持度的候选项集得到频繁K+1-项集；
3. 重复步骤（2）直到无法产生新的频繁K-项集；
4. 挖掘关联规则：通过上述得到的所有频繁项集，取它们的非空子集，则这些子集与去补集形成一种关联关系，如频繁项集{S,T}有S->T和T->S两种关联关系，算出每种规则的置信度，将置信度高的改造成为强关联。
   * 1. 算法优缺点

Apriori算法的优点有：原理简单，实现容易。

该算法的不足之处为：时间复杂度较大，消耗的时间和性能较大，计算置信度和支持度时需要遍历所有的样例数据，而且计算的次数很多。

* 1. 研究方法
     1. 数据集下载及预处理

本研究的最终目的是设计一个网络异常事件检测系统，系统核心算法的编写测试以及系统的功能测试都需要大量的数据进行训练和测试。本研究数据均从网上下载，数据来源于多个互联网公司，经过脱敏处理后供研究者使用。

由于数据均为实时记录的线上数据，因为数据存在一些缺失值需要进行预处理方能使用，本研究采用matlab的插值函数进行数据补全，该函数通过插值点最邻近的两个点的线性函数进行预测。

* + 1. Apriori算法设计与实现

其中数据训练集每个样例包含四个属性，时间戳（timestamp）、值（value）、标签（label）和KPI ID，其中需要挖掘出KPI ID、value和label之间的关联关系，即找寻异常标签的频繁模式，且这些标签在正常模式中不频繁出现。

本研究结合Apriori算法，具体实现如下：

* 1. 实验结果及结论
  2. 本章小结

本章介绍了与本设计相关的数学算法，包括算法的原理、相关知识、不同的实现算法的介绍和对比，并且介绍了实现的过程，包括数据的下载和预处理、算法的完成过程。最后阐述算法的运行的结果，总结实验的过程和收获，给出结论。

1. 网络异常事件检测系统的设计和实现
   1. 需求分析

本研究为设计一个基于频繁模式挖掘的网络异常事件检测系统，通过分析和调研，发现系统最好可以数据可视化、结果可视化和选择不同类型的KPI进行检测等功能，按照系统的定位和调研结果，得出了以下的需求分析。

首先用户可以通过系统选择需要检测的KPI数据对应的训练集和测试集。为满足用户的需求，本系统对KPI的检测不限定于某些类型，而是可以选择多种不同类型的KPI进行检测。

选择了数据文件后，系统会读取文件的内容，并识别出其中包含的KPI种类，以列表的形式将所有的KPI ID展示给用户，用户可以通过绘图功能选择指定的KPI，将其以时间戳为横坐标，value为纵坐标绘出曲线图，并标出其中的异常点，并计算该KPI中异常所占的比例，以饼图形式展示。此功能可以让用户更加直观的查看KPI的模式，以及查看异常发生的情况和比例。

最后还有一个十分关键的功能，即对不同种类的KPI实行测试。用户可以通过检测功能对读取的测试数据的KPI进行检测，用户可以选择不同类型的KPI进行检测。系统检测完成后，同样会将测试数据的KPI曲线图绘画出来，同时展示原始异常和算法预测异常的对比，计算系统预测的精确率和召回率，以可视化的形式展示在用户面前。

* 1. 系统功能设计

系统功能图如下所示：

用户界面

功能接口

读文件功能

绘图功能

检测功能

导出文件功能

文件读取服务

数据解析服务

数据存储服务

绘图服务

异常计算服务

异常模式挖掘服务

测试集检测服务

文件导出服务

图1 系统功能设计图

* + 1. 读取文件

用户点击文件按钮即可选择测试集和训练集，待系统读取文件完毕后，系统将用户选择的文件显示在界面。

* + 1. 绘画KPI曲线图

系统读取用户数据后，会从训练集中找出KPI ID的种类，并将ID的列表展示在界面，用户可以选择想要查看的KPI，点击绘图功能，即可将该KPI的曲线图绘制在界面上，并且会标出其中的异常点，同时以饼图的方式展示异常所占的比例。

* + 1. 异常频繁模式挖掘

选择文件之后，用户即可使用系统对训练集训练，只需选择训练功能，系统会自动开始挖掘异常标签中的频繁模式。

* + 1. 测试及评估算法

进过训练数据后，得到异常标签的频繁模式，用户即可使用挖掘的模式对测试集进行测试，系统测试完毕后，同样会将测试的异常点以曲线图的形式展示给用户，同时展示原始的数据异常图便于用户对比。并算出测试成果的精度、召回率和得分。

* + 1. 导出测试集

对测试集检测结束后，系统展示检测的结果和得分，用户可将测试的结果导出到文件中，以便后续观结果时可以直接观察文件。

* 1. 数据流图

用户导入的数据流转图如下所示：

用户

训练集和测试集

KPI ID

KPI数据文件

KPI ID

KPI数据

数据图形

KPI ID

KPI数据

检测结果

图2 系统数据流图

* 1. 系统功能实现
     1. 系统界面设计

按照系统的需求分析和功能设计，系统界面以简洁实用为原则，在保证所有需求实现的同时，给用户带来良好的实用体验。

图图图图图

上图为系统的外观界面图。该界面分为三个模块，分别为顶端的菜单栏、中部的文件选择和绘图模块以及底部的KPI检测模块。

菜单栏为用户展示的的操作包括文件、编辑、工具和帮助四个操作。每个操作的功能如下：

1. 文件菜单提供了打开文件、保存文件和退出程序的功能；
2. 编辑菜单提供了清除数据、选择曲线图颜色和缩放图片的功能；
3. 工具菜单栏提供了图片下载、测试结果导出功能；
4. 帮助菜单栏提供了软件介绍和使用帮助的功能。

中部文件选择和绘图模块主要的功能为：

1. 可用通过菜单栏或者文件选择按钮选择需要读取的训练集和测试集；
2. 文件选择完毕后，系统会展示训练集和测试集的文件名以便用户查看；用户选选择文件结束后，软件会读入文件中的数字和文本，并将这些数字和文本存储到软件的工作空间中，为其他功能提供数据。
   * 1. 系统后台实现

上述章节已经对系统的需求分析、功能设计和界面设计进行了阐述，接下来将分别讲述各个功能的后台实现方式。

1. 读取文件：读取文件使用的是matlab GUI文件选择函数uigetfile()函数，该函数能够提供一个文件选择窗口供用户选择文件，并将用户选择的文件名保存下来。获得用户选择的文件名后，即可使用xlsread()函数读取文件，并将文件数据存储到矩阵中。
2. KPI种类解析：读取文件完毕后，对数据进行解析，将其中的数据和文本分别存储，数据主要是测试集和训练集的KPI 数据，文本则是KPI ID。将KPI ID种类解析后展示在用户面前，解析方法是遍历所有的ID，然后取出重复ID，将ID的种类解析出来。
3. KPI绘图：用户选择某个KPI后，系统查找该KPI所有的数据，然后通过matlab的plot()函数将这些数据绘出曲线图展示给用户，并计算其中的异常比例，使用pie()函数绘出饼图。
4. 异常模式挖掘：通过前面阐述的2.3.2章节中Apriori算法的设计与实现中的方法，挖掘KPI数据中的异常标签的频繁模式。
5. 测试数据异常检测：用户选择需要检测的KPI ID后，通过（4）中挖掘出来的频繁模式判断测试样例是否异常。
6. 召回率和精确率的计算：精确率=TP / (TP + FP)，召回率=TP / (TP + FN)。其中符号的含义是：对于一段标记的连续异常区间：如果算法成功检测到整段异常区间，区间内的每个异常点算作一次TP；否则该连续异常区间内的每个异常点算作一次FN。对于一个没有标记异常的时间点：如果异常检测算法标记异常，记做一次FP，否则记做一次TN。
   1. 设计结果

本设计的结果为成功设计一个基于频繁模式挖掘的网络异常事件检测系统，其中对系统的需求分析的功能全部实现，系统的界面如下：

。。。。。。兔兔兔兔兔兔。。。。

* 1. 本章小结

本章主要介绍了网络异常事件检测系统的设计和实现过程。

首先介绍了软件的需求分析，从用户的角度和软件功能的角度对软件的需求进行了说明，叙述了软件功能和界面设计的理念和思想。

紧接着介绍系统的功能设计，包括系统的主要功能和功能的作用，帮助读者更好的熟悉系统的功能要点。

然后给出了系统的数据流图，描述系统数据的流动方向，帮助读者更好的理解系统处理数据的流程。

接着叙述了软件的实现历程，囊括软件外观的制定和软件后端实现的重要技术和历程，为用户现实界面的重要模块，以及界面与后端共同工作的的方式，让读者了解软件的全部设计心路历程和重要模块。

最后为读者呈现系统的最终形态，即本设计工作最终的结果。

1. 总结与展望
   1. 总结

在互联网时代背景下，保证一个web服务的稳定运行十分重要，仅靠人力的监控和检查是很难远远不够的，如今数据挖掘和分析技术发展的十分迅猛，在一些方面的应用已经十分成熟，所以结合数据挖掘技术设计更加高效的网络异常检测算法很有必要。本文的首要工作包罗如下两个方面：

1. 频繁模式挖掘算法

本文研究频繁模式挖掘算法在网络异常检测中的应用，通过挖掘异常标签的频繁模式，用这些频繁模式对数据进行检测。作者使用的是经典的Apriori算法进行频繁模式挖掘，进实验该算法能挖掘出各种类型KPI的异常频繁模式，并能对测试数据进行检验。

1. 网络异常检测系统设计

本文的另一个重要的模块为制定和完成一个网络异常检测系统，文章中对软件的需求阐明、主要功能、界面设计和实现历程进行了叙述，将制定和实现的全部历程归纳综合的显现给读者。

* 1. 展望

本毕业设计结合频繁模式挖掘算法最终设计了一个功能完善的网络异常事件检测系统，为更多的研究者提供了参考。但是由于能力和时间的局限性，系统的精确度还带提高，后来者可以采用其他算法对Apriori算法进行优化，或者研究更加高效优秀的算法用于异常检测中。

参考文献：

1. 郭通. 基于自适应流抽样测量的网络异常检测技术研究[D]. 中国人民解放军信息工程大学, 解放军信息工程大学, 2013.