第二章 大数据审计分析的技术与工具

新道
seentao

目录
CONTENT

第一节 大数据审计分析技术概述

第二节 大数据审计分析工具概述

新道
seentao

第一节

大数据审计分析技术概述

数据收集技术

新道 seentao

• （1）API

- API是应用程序接口（Application Programming Interface）的简称。API是一些功能、定义或者协议的集合，通过API接口可以实现计算机软件之间的相互通信。API提供应用程序或者程序开发人员基于软件访问一组例程的能力，对外封装完善，调用时无需学习API内部源码，依据API文档功能说明书来使用即可。

• （2）爬虫

- 爬虫即网络爬虫，是指能够自动访问互联网并将网站内容下载下来的程序。爬虫会按照一定的规则自动浏览、检索网页信息的程序或者脚本，它能够自动请求网页，并将所需要的数据抓取下来。通过对抓取的数据进行处理，从而提取出有价值的信息。

• （3）预定义规则处理

- 预定义规则处理是指把执行的语句编译成计算机能够理解的形式，主要过程有数据抽取（Extraction），数据转换（Transformation）和数据加载（Loading），也称为ETL，这个过程是负责将分布的、异构数据源中的数据抽取等到临时中间层进行转换、集成等处理，最后加载到数据仓库或数据集市中。

数据清洗技术

新道 seentao

- **（1）结构化**

- 结构化是指对采集到的数据在分析之前将非结构化数据转换为结构化数据的过程。大数据技术擅长在一定规则下对大量有规律的结构化数据进行建模处理，如果直接使用半结构化或非结构化数据进行数据分析则难以得到理想效果

- **（2）标准化**

- 标准化是指通过一定的数学变换方式，将原始数据按照一定的比例进行转换，使之落入到一个小的特定区间。数据标准化处理主要包括指标一致化处理和无量纲化处理两种类型。前者主要解决的是数据之间不同性质的问题，后者主要解决数据之间可比性的问题。

- **（3）模糊匹配**

- 模糊匹配是用于比较两个或多个记录并计算它们属于同一主体的可能性。模糊匹配不是将记录大致分类为匹配和不匹配，而是输出一个数字（通常在**0-100**之间），用于标识这些记录属于同一主体的可能性。

数据分析技术

（1）回归分析

在统计学中，回归分析指的是确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法在大数据分析中，回归分析是一种预测性的建模技术，它研究的是因变量（目标）和自变量（预测器）之间的关系。常见的回归方法包括普通最小二乘回归、对数几率回归、多元自适应回归、局部散点平滑估计回归等等。

（2）聚类分析

聚类分析是指将数据对象的集合分组为由类似的对象组成的多个类的分析过程。聚类就是一种寻找数据之间内在结构的技术，它把全体数据实例组织成一些相似组，而这些相似组被称作簇。聚类使组内样本差异极小化，组间样本差异极大化。常见的聚类方法包括K值聚类、层次聚类、模糊聚类、单连锁聚类、期望最大值聚类、非负矩阵分解聚类等等。

（3）文本分析

文本分析是将非结构化文本数据转换为有意义的数据进行分析的过程，以度量客户意见、产品评论、反馈，提供搜索工具、情感分析和实体建模，从而支持基于事实的决策制定。

（4）关联规制学习

关联规则学习又叫关联分析，即从大规模数据集中寻找物品间隐含的关系。关联规则用来描述两个或多个事物之间的关联性，其通过一件或多件事物来预测其它事物，可以从大量数据中获取有价值数据之间的联系。常见的关联规制学习算法包括Apriori算法、Eclat算法、FP-Growth算法等等。

（5）降维

降维是将高维数据集转换为可比较的低维空间的过程，真实的数据集通常有很多冗余特征，降维技术可用于去除这些冗余特征或将n维数据集转换为2维或3维进行可视化。常见的降维技术包括主成分分析、因子分析、判别分析、局部线性嵌入、Sammon映射、投影寻踪等等。

（6）集成方法

集成方法是指通过将一系列相对较弱的模型以某种恰当的方式组合起来，可以得到比单个模型效果更好的强模型，从而提高模型的性能。集成方法可以很容易地减少过拟合，避免模型在训练时表现更好，而在测试时不能产生良好的结果。常见的集成方法包括Boosting、自展输入引导式聚合、Adaboost、堆栈泛化、随机森林等等。

（7）决策树

决策树是一种树形结构（例如二叉树），其中每个内部节点表示一个属性上的判断，每个分支代表一个判断结果的输出，最后每个叶节点代表一种分类结果。常见的决策树算法包括分类回归数、迭代二叉树、卡方自动交互检测、单层决策树、条件决策树等等。

（8）贝叶斯

贝叶斯是利用概率统计知识进行分类的算法统称。常见的贝叶斯算法包括朴素贝叶斯、高斯朴素贝叶斯、多项式朴素贝叶斯、平均单依赖分类器、贝叶斯信念网络、隐马尔可夫模型、条件随机场等等。

（9）神经网络

神经网络是由大量简单处理单元按不同方式互相连接构成的并行分布式信息处理系统，这些处理单元也被称为神经元、神经节点。它模仿人脑神经系统，通过对预先提供的一批相互对应的输入输出信号进行学习分析，挖掘出两者之间的潜在规律，然后根据这些规律完成对新输入信号推算出输出结果的处理。常见的神经网络算法包括自组织映射、感知机、反向传播算法、霍普菲尔德神经网络、径向机函数网络、玻尔兹曼机、受限波尔兹曼机、Spiking神经网络、学习矢量量化等等。

（10）深度学习

深度学习是利用机器学习算法让模型学习数据的内在规律和表示层次，通过学习过程帮助机器获得对诸如文字、图像、声音等数据的解释。常见的深度学习算法包括深度玻尔兹曼机、深度信念网络、卷积神经网络、堆栈式自动编码器等等。

1. 多层感知机(Multilayer Perceptron，MLP):一种基本的前馈神经网络模型，用于解决各种机器学习和深度学习任务，如分类、回归和模式识别等。

2.卷积神经网络(Convolutional Neural Networks，CNN):主要用于图像处理和计算机视觉任务，通过卷积层和池化层来提取图像中的特征。

3.循环神经网络(Recurrent Neural Networks，RNN):用于序列数据建模，具有循环连接的结构，可以处理时间序列数据、自然语言处理等任务。

4.长短时记忆网络(Long Short-Term Memory,LSTM):一种特殊类型的RNN，用于解决传统RNN中的梯度消失问题，适用于处理长序列。

5.门控循环单元(Gated Recurrent Unit，GRU):与LSTM类似，但具有更少的门控单元，参数较少，适用于一些轻量级的序列任务等等。

数据可视化技术

（1）直方图

直方图是一种统计报告图，用来展示数据值分布情况的图形。一般横轴表示数据区间，纵轴表示数据分布。图2-2展示了一个反映现金支出分布的直方图图例。

（2）散点图

散点图通常用于比较跨类别的聚合数据，展示数据点在直角坐标系平面上的分布，判断变量之间是否存在某种关联或总结坐标点的分布模式。散点图将序列显示为一组点，值由点在图表中的位置表示，类别由图表中的不同标记表示。图2-3展示了一个反映渠道销售订单分布的散点图图例。

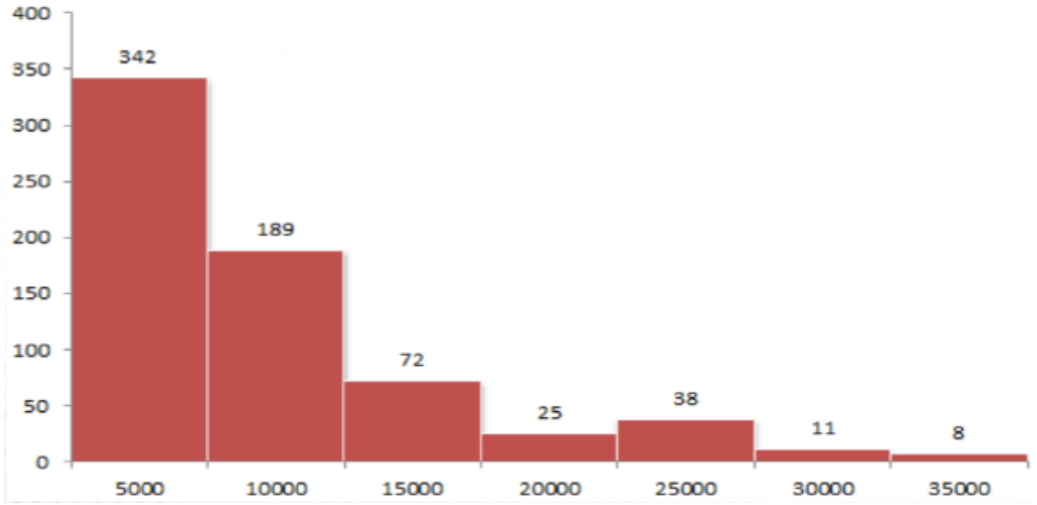


图2-2 直方图图例

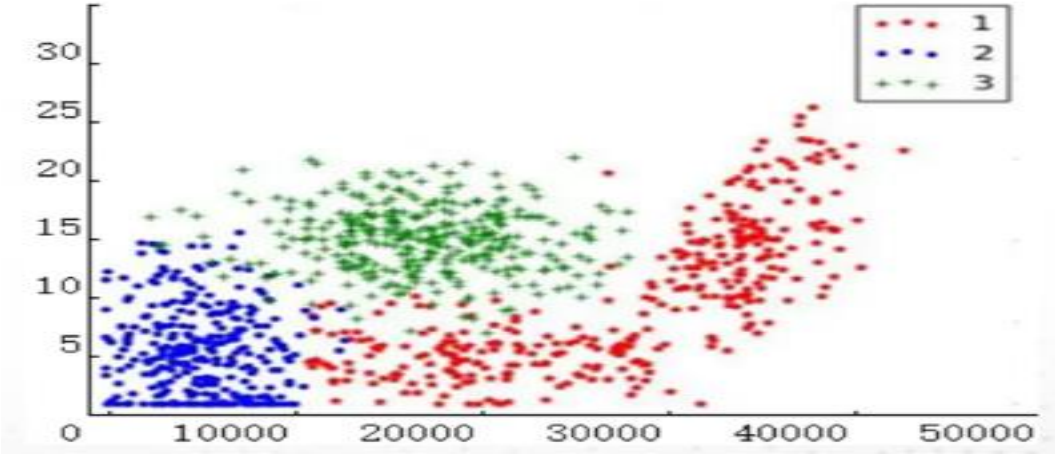


图2-3 散点图图例

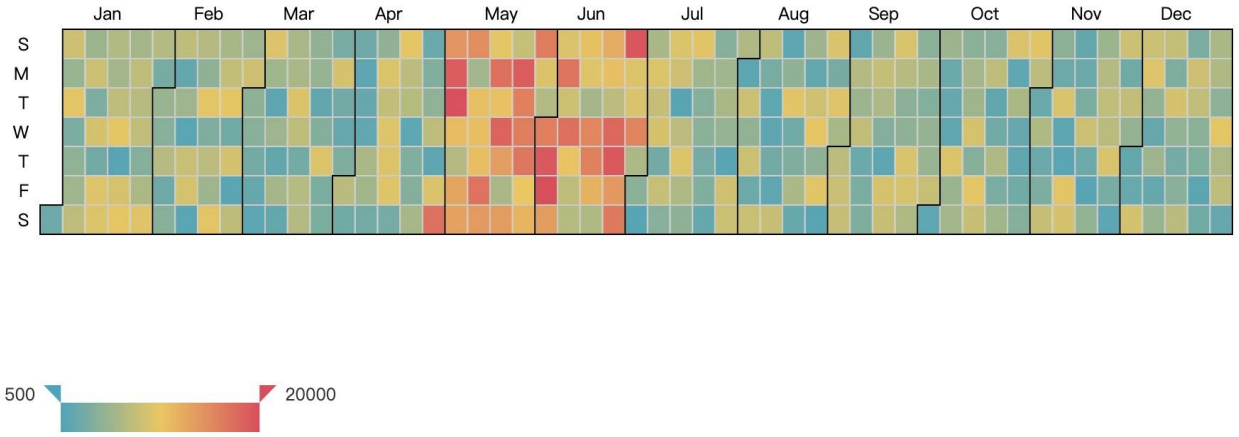
数据可视化技术

（3）K线图

K线图又叫蜡烛图。从K线图中，既可以看到股价的趋势，也可以了解到每日股价的波动情况，能够全面透彻地观察到市场价格的真正变化。K线图主要显示值为最高价、最低价、开盘价和收盘价。左图展示了一个反映股价日趋势的K线图图例。

（4）热力图

热力图是指根据所属部分的颜色深浅来反映数值大小的图形。热力图可以直观地将数据分布通过不同颜色区块呈现。热力图的一种典型应用是以地图的形式展现数据分布，同时还可以显示对应地理位置。日期热力图也是比较常见的热力图应用，可以反映不同日期的数据量情况，右图展示了一个年度内关于某上市公司的讨论量情况，可以看到在5月、6月年报公布后，讨论量有明显上升趋势。



数据可视化技术

（5）词云图

词云就是通过形成关键词云层或关键词渲染，对文本中出现频率较高的关键词在视觉上进行突出，关键词出现频率越高，面积越大。词云图过滤掉大量的低频低质的文本信息，使浏览者只有一眼扫过文本就可以领略文本的主旨。图2-6展示了一个反映股吧股民评论的词云图图例。

（6）社会网络图

社会网络图是一种基于网络（节点之间的相互连接）的可视化展现形式，可以直观的看到各个主体在网络中的位置和网络整体结构，用于挖掘分析主体之间的社会网络关系。

图2-6 词云图图例

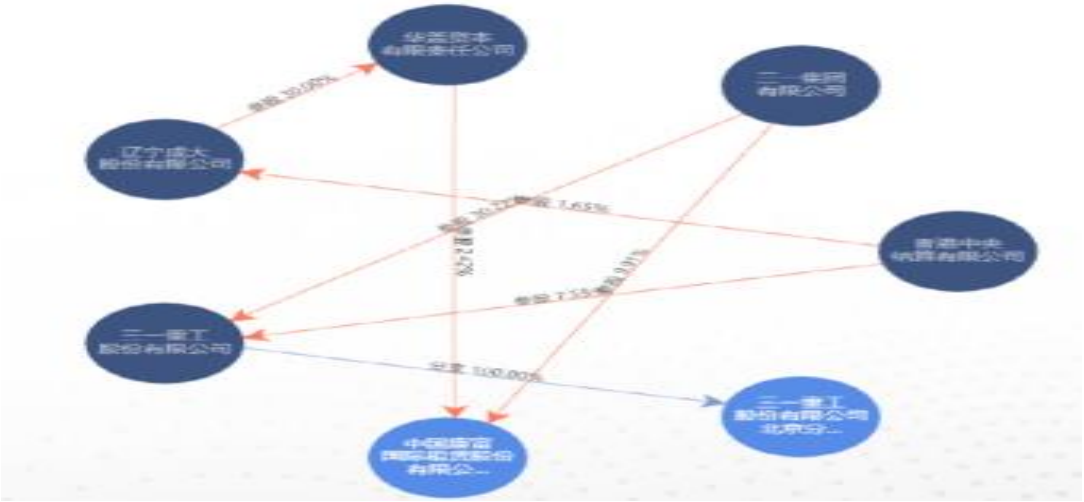


图2-7 社会网络图图例

第二节

大数据审计分析工具概述

大数据审计分析工具概述

考虑到大数据审计分析应该满足对被审计单位各部门、各环节多种类型全样本数据的使用需求，而不仅仅局限于特点的财务、业务数据，且对数据的使用需要打破时间和空间的限制，满足随时随地进行多维度审计的需求。因此结合配套的实践平台，本书介绍了三个非常常见，且能够满足大数据审计分析应用需求的具体工具。

SQL基础

Python基础

RPA基础

SQL基础

（1）SQL简介

SQL（Structured Query Language，结构化查询语言）是一种计算机标准语言，通过SQL我们可以对数据库执行查询、新增、更新、删除等操作。SQL是对埃德加·科德的关系模型的第一个商业化实现，尽管SQL并非完全按照科德的关系模型设计，但其依然成为最为广泛应用的数据库语言。SQL在1986年成为美国国家标准学会（ANSI）的一项标准，在1987年成为国际标准化组织（ISO）标准。SQL语句主要是与关系数据库管理系统进行数据交互，常见的关系数据库管理系统有：MySQL、Oracle、DB2和SQLServer。每个数据库管理系统的SQL语言风格略有不同，但他们的操作都很相似，而且都是基于标准SQL规范的。

（2）SQL在大数据审计分析中的应用

随着经济的高速发展，企业经营数据暴增，在审计工作中经常会遇到数据量很大的情况，比如一个工作簿占用的内存有30多M，一个工作表里面有30多万行数据。审计人员遇到这种情况，常常感觉自己电脑反应很慢，即使高配的电脑也是如此。这个时候，使用SQL就可以进行批量化的数据操作。

SQL基础

（3）SQL基本语法

1·不区分大小写： SQL语句不区分大小写，意思是我们可以使用大写或者小写，效果都是一样的。但是为了便于阅读和调试代码，比较规范的做法是，SQL的关键字大写，表名或者字段名小写。

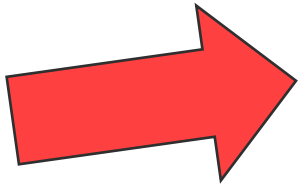
2·多条语句分号分割： 对于单条SQL语句来说，在结尾处加分号或者不加分号都是可以的，但是多条SQL语句必须以分号分隔。

3·必须英文标点符号： SQL语句中所使用的标点符号，需要是英文状态的标点符号，如果使用中文的标点符号会报错。

4·空格会被忽略： 在处理SQL语句时，所有的空格都会被忽略，我们可以把一条SQL语句写在一行上，也可以分开写在多行上。比较好的习惯是将SQL语句写在多行上，这样使得代码更容易阅读和调试。

表2-1 SQL添加注释规则

如右图



注释情形	示例
和SQL语句在同一行：用“- -”	SELECT prod_name FROM products; - -这是一条注释
单独一行：用“#”	#这是一条注释SELECT prod_name FROM products;
多行注释：用“/*　*/”	/*这是一条注释 这是一条注释 这是一条注释*/ SELECT prod_name FROM products;

SQL基本语句

常用SQL基本语句如表2-2所示。

基本语句	语句功能	基本语句	语句功能
CREATE DATABASE	创建数据库	SHOW	显示(数据库/数据表)
CREATE TABLE	创建数据表	USE	选择数据库
DROP	删除（数据库/数据表）	SELECT	从数据库中查询数据
DELETE	删除数据库中的数据	SELECT * FROM...	从某表中查询数据
DELETE FROM	从某表中删除数据	SELECT * FROM...WHERE ...	从某表中查询符合某些条件的数据
DELETE FROM...WHERE ...	从某表中删除符合某些条件的数据	SELECT DISTINCT	返回唯一不同的值
UPDATE	更新数库中的数据	SELECT * FROM...ORDERBY……DESC	从某表中查询数据后按照某个字段进行降序排列
INSERT INTO	在数据库中插入新的数据	SELECT COUNT(*)FROM...	查询某表中的数据条数
LOAD DATA	导入数据	Group by	以某字段唯一值为汇总依据汇总数据

SQL基础

新道 seentao

常用SQL基本语句列表

基本语句	语句功能	基本语句	语句功能
CREATE DATABASE	创建数据库	SHOW	显示(数据库/数据表)
CREATE TABLE	创建数据表	USE	选择数据库
DROP	删除（数据库/数据表）	SELECT	从数据库中查询数据
DELETE	删除数据库中的数据	SELECT * FROM...	从某表中查询数据
DELETE FROM	从某表中删除数据	SELECT * FROM...WHERE ...	从某表中查询符合某些条件的数据
DELETE FROM...WHERE ...	从某表中删除符合某些条件的数据	SELECT DISTINCT	返回唯一不同的值
UPDATE	更新数库中的数据	SELECT * FROM...ORDERBY……DESC	从某表中查询数据后按照某个字段进行降序排列
INSERT INTO	在数据库中插入新的数据	SELECT COUNT(*)FROM...	查询某表中的数据条数
LOAD DATA	导入数据	Group by	以某字段唯一值为汇总依据汇总数据

SQL基础

常用SQL基本语句示例如图表2-3至表2-9所示

语句用法	语句示例
SHOWDATABASES：列出MySQL数据库管理系统的数据库列表	#显示MySQL中所有的数据库SHOW databases;
CREATE DATABASE+数据库名：新建数据库	#创建名为my_database的数据库CREATE DATABASE my_database;
DROP DATABASE+数据库名：删除数据库	#删除数据库my_databaseDROP DATABASE my_database;

表2-3 SQL基本语句——增删数据库

表2-4 SQL基本语句——增删数据表

语句用法	语句示例
SHOWTABLES:显示指定数据库的所有表。使用该命令前需要使用USE命令来选择要操作的数据库	#显示py_database数据库中所有的表USE py_database;SHOW tables;
USE+数据库名：选择数据库	
CREATE TABLE+数据表名(“字段名”字段定义)：新建数据表CREATE TABLE table_name(column_name column_type);使用该命令前需要使用USE命令来选择要操作的数据库	#在py_database创建名为my_table的数据表USE py_database;CREATE TABLE my_table(cust_id char(10) NOT NULL,cust_name char(50) NOT NULL);#char(10)以及char(50)指定字段字符的长度#NOTNULL指定字段不能为空值，如不指定，默认为NULL，即可以为空值
DROP TABLE+数据表名：删除数据表。使用该命令前需要使用USE命令来选择要操作的数据库。	#删除数据表my_tableUSE py_database;DROP TABLE my_table;

【增删数据库】

表2-3 SQL基本语句——增删数据库

SQL基础

【查询数据】

表2-5 SQL基本语句——查询数据

语句用法	语句示例
SELECT+字段名：要查什么	#从数据库表products中查询prod_name
FROM+数据表名：从哪里查	**SELECT prod_name FROM products;**
DISINCT+字段名：表示数据库只返回该字段不同的值（查询结果去重）	#从数据库表products中查询不重复的prod_name值SELECT DISTINCT prod_name FROM products;
查询多列时，各个列之间以逗号分隔，最后一列的后面不加逗号	#从数据库表products中查询prod_id和vend_idSELECT prod_id, vend_id FROM products;
查询多列均不重复的值时，同样使用DISTINCT，其作用于所有的列，不仅仅是跟在其后的一列，也就是所有列值都相同的行才会被去除，并且DISTINCT要放在所有要查询的列的最前面，不可以放在中间位置	#从数据库表products中查询vend_id和 prod_price均不重复的数据行SELECTDISTINCT vend_id, prod_price FROMproducts
查询所有列时，可以列示所有列名，或者使用*号通配符来实现	#从数据库表 products 中查询所有列SELECT * FROM products;

注：使用表2-5所示命令前需要使用USE命令指定数据库。

WHERE语句可以指定查询条件，可以在WHERE子句中使用的操作符包括：=，>，<，>=，<=，<>，BETWEEN AND，LIKE等。

其中，LIKE操作符，主要在过滤模糊值时使用。一般会与通配符（%、_）结合使用。

%为通配符，表示任何字符出现任意次数。与%能匹配若干个字符不同，_总是刚好匹配一个字符。通配符查询只能用于文本字段，非文本数据类型字段不能使用通配符搜索。

多条件查询时，可以使用OR、AND操作符将多个条件组合在一起。AND 用来指示查询满足所有给定条件的行。OR用来指示查询满足任一给定条件的行。