上海交通大學

《数据可视化与可视分析》课程设计论文

课设题目：ChinaVis 2018 挑战 1 的可视化分析方案与设计

学生姓名：强志文　　学号：515030910367

学院（系）：计算机科学与技术系

同组同学：张浩诚，熊俊，曹以想

2018 年 6 月

ChinaVis 2018 挑战 1 的可视化分析方案与设计

1. 作者信息

姓名：强志文
学院：电子信息与电气工程学院
班级：F1503015

2. 正文

2.1 问题描述

某互联网公司有几百名员工，分属财务、人力资源和研发三个部门。公司正在全力研发一款重量级新产品，近期该产品临近发布，公司对内部发生的一切异常现象都非常敏感。为了维护公司的核心利益，确保新产品顺利发布，公司高层决定临时成立内部威胁情报分析小组，该小组将根据公司内部采集到的数据，分析并处置可能存在的各种安全威胁。假设您是威胁情报分析小组的成员，请您设计并实现一套可视分析解决方案，帮助该公司及时准确地找出可能存在的内部威胁情报[1]。

（1）分析公司内部员工所属部门及各部门的人员组织结构，给出公司员工的组织结构图

（2）分析该公司员工的日常工作行为，按部门总结员工的正常工作模式

（3）找出至少 5 个异常事件，并分析这些事件之间可能存在的关联，总结你认为有价值的威胁情报

2.2 设计思想

首先我们需要一个视图反映员工的从属情况，他们属于什么部门，各个部门的人员组织结构如何。于是我们想到了用力导向图来反映这些情况。每个节点代表员工，节点与节点之间的连线表示他们之间有群发邮件的往来。在这里之所以选择群发邮件信息是因为我们认为这一信息最能反映部门与部门之间的关系。如果单纯的以邮件联系作为标准，则其中可能会掺杂许多无用的私人信息。于是我们确立了视图 1。

同时所有的数据都有相应的天数数据，我们想到了使用一个时间轴模块来表示天数信息。这样可以很好的表示数据的时间特征。同时时间轴模块还会反应员工的上班情况，也是提供给了用户更多的信息。于是我们确立了视图 5。

再者我们需要了解一个群体的数据信息，这样有助于在群体中发现异常的个人，同时可以有效的发现群体基本数据情况。于是我们确定了视图 3。

我们在通过视图 3 确定了可能存在异常的个人后，需要更加细致的观察该员工的具体信息。于是我们有了视图 2，可以表示选定的员工在一定天数内的上下班时间，上行，下行流量信息。同时也确立了视图 4，可以反映选定的个人在 30 天内每个小时的各种协议的流量信息。最后还有视图 7，可以反映选定的员工在某一天内的各种协议信息的综合情况。

最后，我们通过群发邮件的信息可以将各个部门区分开，但是我们无法确定每个群体到底属于哪个部门，于是我们设计了视图 6，通过反映一个群体在一段时间内群发邮件的主题信息，我们可以确定群体所属部门类别（财务、人力资源还是研发），同时让观察者对于部门情况有一个更深入的了解。

2.2 系统总体设计方案(包括交互和联动)

系统如图 1 所示，共分为 7 个模块: 员工分布模块(视图 1)、个人流量展示模块 (视图 2)、群体流量分布模块 (视图 3)、个人流量时间分布模块 (视图 4)、时间轴轴模块 (视图 5)、群体邮件信息模块 (视图 6)和个人单日流量信息汇总模块 (视图 7)。

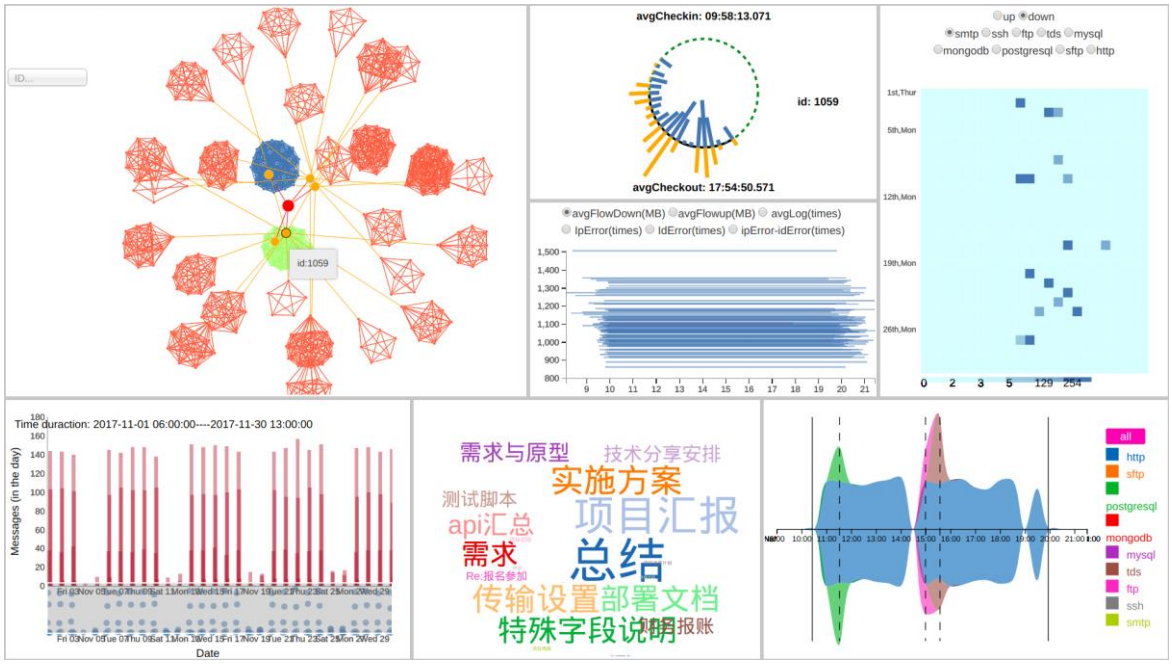

图 1 系统界面

员工分布模块(左一上一): 通过员工之间的群发邮件建立力导向图展示了该互联网公司的员工分布。个人流量展示模块 (左二上一): 展示了某位被选中员工在一段时间内产生的总上下行流量的时间分布和平均上下班时间。群体流量展示模块(左二上二): 展示了某位被选中的员工的群体在一段时间内的总上下行流量和各 id，ip 的登录错误次数统计。个人流量时间分布模块(左三上): 展现了某位员工一个月以小时为单位每小时的某种流量的时间分布。时间轴模块(左一下一): 展示了一个月内员工出勤的时间分布。群体邮件信息模块(左二下一): 展示了某位被选中的员工的群体在一段时间内的邮件主题信息。个人单日流量信息汇总模块(左三下一): 展示了某位员某一天内所有流量的汇总图及登入登出账号和上下班情况。

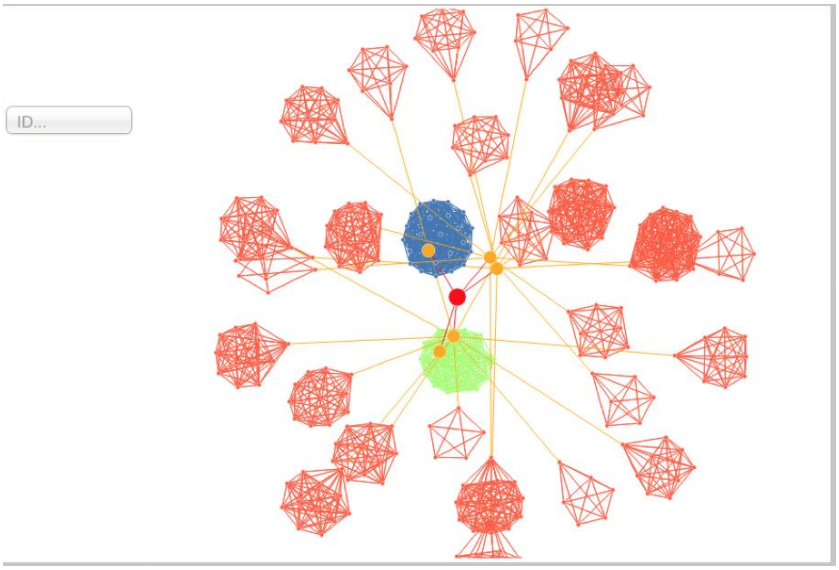

图 2 员工分布模块(视图 1)

图 2 中展示的是员工分布情况(视图 1)，可以看出红点所代表的员工为公司的中心(总裁)。黄点所代表的员工为每个部门中向总裁汇报情况的部门领导。其他点则分为许多个小群体。其中蓝色代表财务部门，绿色代表人力部门，橙色代表研发部门。研发部门还分为许多个小组，每个小组有与部门领导交流的二级领导。

在视图 1 中我们可以选中某个点的获取该点的 id 并联动视图 2，3，4，6 使其展示该 id 的相应数据。也可以在矩形框中直接输入想要探索的 id 进行联动。

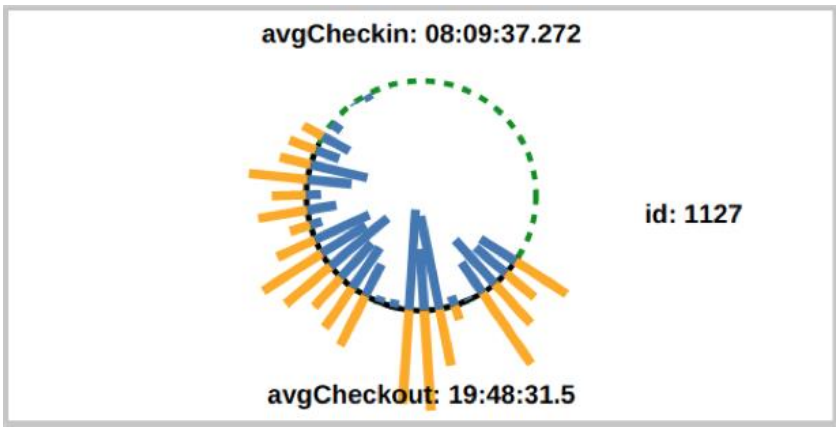

图 3 个人流量展示模块（视图 2）

图 3 中展示的是单个员工在一段时间内的流量情况(视图 2)。输入为视图 1 中选中的员工和视图 5 中选中的时间段。

其中员工的平均上下班时间以文本的形式展现在视图中。而员工的总上下行流量的时间分布则展示为柱形图在圆上的分布，其中红色柱形图表示上行流量，蓝色柱形图表示下行流量。

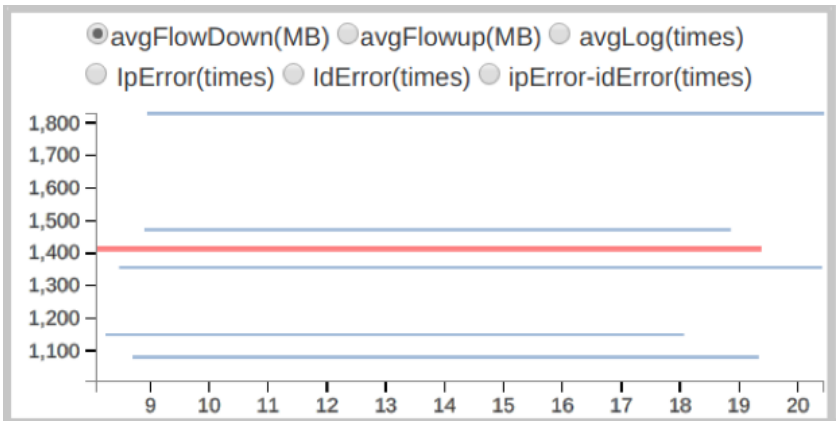

图 4 群体流量分布模块（视图 3）

图 4 中展示的是某个员工群体在一段时间内的流量分布情况(视图 3)。输入为视图 1 中所选中的员工所在的群体和视图 5 中选中的时间段。

其中每一条横线代表一位在该群体中的员工，横线的 y 轴为所观察流量的大小，横线的 x 轴的两端分别为该员工的平均上下班时间。可被观察的流量有：总下行流量，总上行流量，总登陆次数，ip 登录失败次数，id 被登录失败次数，ip 登录失败次数与 id 被登录失败次数之差

当鼠标移到某一个员工的横线上时会显示出该员工的 id。

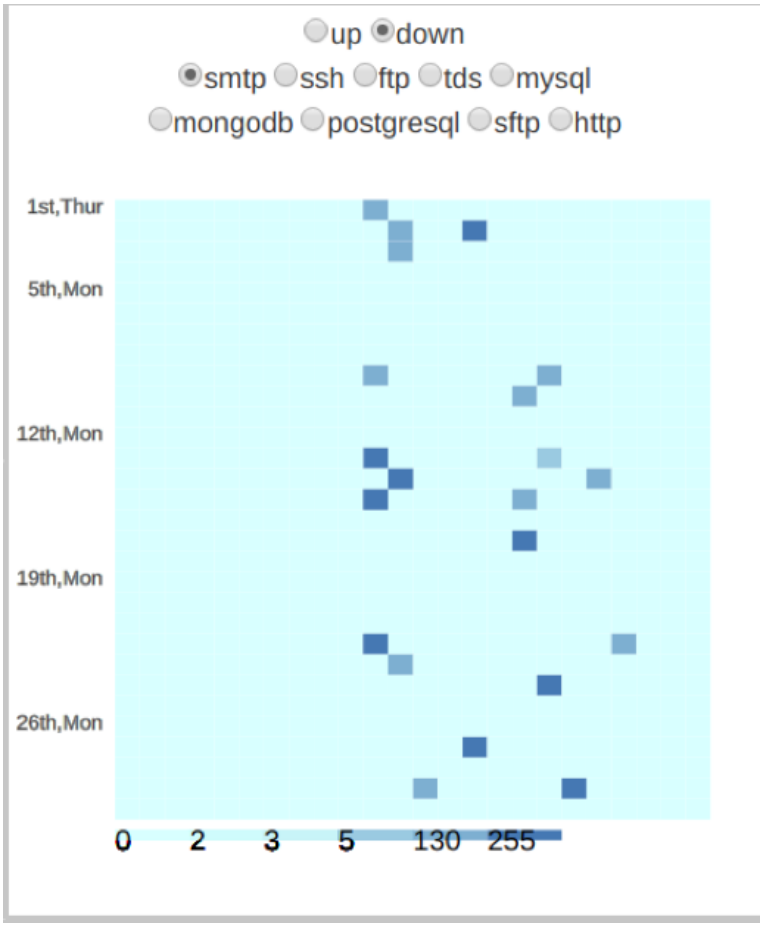

图 5 个人流量时间分布模块（视图 4）

图 5 中展示的是某个员工在一个月内各个流量的时间分布(视图 4)。数据输入为视图 1 中所选中的员工。

图中的每一个矩形框代表一个小时，每一行有 24 个小时代表一天。二矩形框的颜色深浅泽表示该员工需要被观察流量的大小。在筛选框中，可以选择任意协议的上行或下行流量。

当鼠标悬浮在某一矩形框中时，会展示出该矩形框所代表的具体时间和流量值。当鼠标点击某一矩形框时，会把相应的日期传递给视图 6 进行联动。

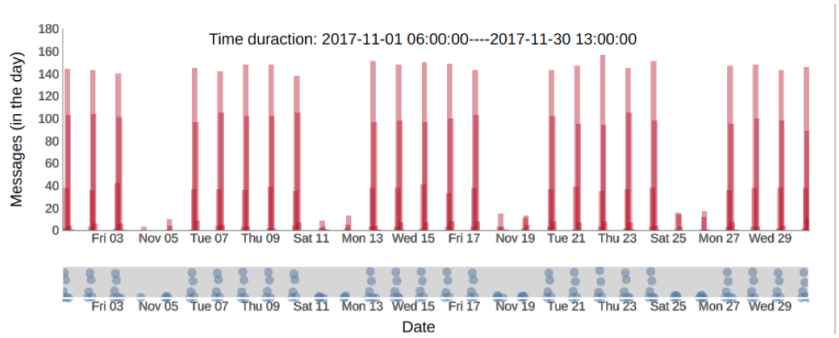

图 6 时间轴轴模块（视图 5）

图 6 中展示的是一个月内所有员工的出勤时间分布(视图 5)。在下方的时间轴中可以通过矩形框选中某一段时间并在上方放大，展示为矩形图。该矩形图的时间精度精确为每半小时。

当下方时间轴的矩形框选中某一段时间后，该时间段会被传递给视图 2，3 并进行联动。而当鼠标悬浮在上方的矩形图中某一矩形时，会展示出该矩形所统计的所有员工的 id，并在在视图 1 中进行标记。

图 7 群体邮件信息模块（视图 6）

图 7 展示了某位被选中的员工的群体在一段时间内的邮件主题信息（视图 6）。输入为视图 1 中所选中的员工所在的群体和视图 5 中选中的时间段。

其中文字的大小表示该内容在该群体邮件中提及的频率，字越大，则该内容在邮件主题中出现的次数越多，相应的也越重要。图 7 为某研发部门群体的邮件信息，可以看出主要内容为“方案”，“脚本”等，这也印证了我们在视图 1 中的群体分类的正确性。

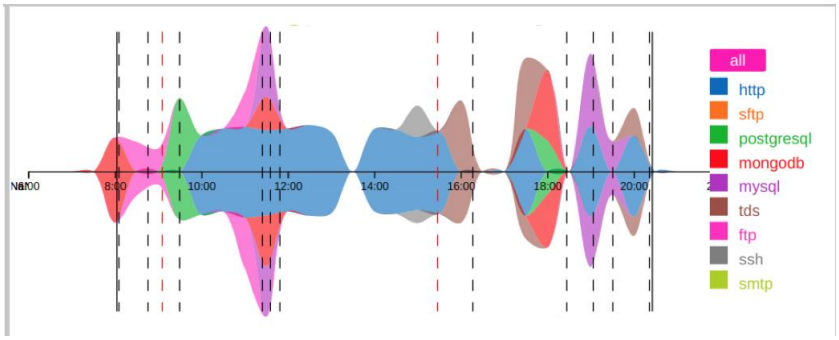

图 8 个人单日流量信息汇总模块（视图 7）

图 8 中展示的是某个 id 的员工某天所有流量情况的汇总(视图 7)。该视图的输入为视图 1 中选择员工 id，和视图 4 中所选择的日期。

图中每个颜色都代表着某种协议的流量，用连续的曲线图展现了它们一天内的时间分布。而该员工该日的上下半时间则由实线表示，登录登出账号由虚线表示。若登录失败，则虚线为红色。

2.4 自己设计部分的主要实现

2.4.1 视图 1，视图 2，视图 3，视图 6 的设计与实现

对于视图 1，力导向布局算法的实现我使用了 d3.js 自带的库函数，通过接受后端传来的员工分组信息绘制而成。同时我调节了算法内部的参数，使群体内部的节点相对紧密，群体与群体之间的节点距离相对较远。

对于视图 2，展示了单个员工在一段时间内的流量情况，首先我绘制了一个圆环，圆环的弧度反应了员工的上下班时间，环内侧和外侧的圆柱分别反应了员工的下行流量和上行流量。对于这一部分的实现主要就是接受后端的数据，绘制基本的形状即可。

对于视图 3，展示了一个群体的上下班时间，流量和登录错误信息；对于这一部分主要一个单选框和折线图。主要就是分清不同数据的特点，首先确定最大值和最小值，而后，每条横线代表一个员工，横线的起始位置和终止位置代表其上班时间和下班时间，用户通过选择单选框某个属性，而后横线的高度即代表该属性（流量，登录错误等）在整体的表现情况。通过这一视图，用户可以对于其所选择的群体的基本信息有一个大致的了解。

对于视图 6，展示选中的群体的邮件主题信息，我借鉴了关于文字云的库函数，该函数接受一定数量的文字和对应文字出现的次数，而后就可以将文字按一定规律排布出来。一般文字越大，代表该文字出现次数越多，相应的也越重要。比如在研发部门中，主要内容为“接口”，“技术”等等，在财务部门为“报账”等等。该视图可以让用户对于每一个群体的工作有一个大致的了解，同时也印证了视图 1 中不同群体的分类的正确性。

2.4.2 系统的交互和联动

整个系统的交互联动都是我设计并实现的。交互和联动在 2.3 中已有说明，在这里不一一举例。该部分的难点在视图 2，视图 3，视图 6 需要需要根据视图 1 员工的选择和视图 5 时间区域的选择而变化。所以我将时间信息和员工节点信息作为这三个视图的私有变量保存，有变化便及时更新。

2.4.3 协助另外两位同学完成他们的视图部分

在小组讨论中，我完善并确立了视图 4 的布局，视图 5 的布局也是我提出来的，并由其他组员加以实现。除此之外，我也帮助其余两位同学确立了代码风格和规范。

2.5 结论

2.5.1 研究工作总结

通过前期对数据的分析可以初步了解数据的分布特性，再结合需要解决的问题以及待处理的难点的后对数据进行重新筛选、统计、分类等预处理。接着以不同的角度和不同的层次对数据进行可视化，能对数据进行更为全面深入的探索，而且利用视图模块间的交互加强系统的整体性和层次性，将不同层次、不同属性的信息联系在一起，发现数据背后的故事，这也是我们进行可视分析的目的。

最终，我们的可视化设计方案包含 7 个视图，彼此之间互有关联，层层递进，借助它们视图本身以及其间的联动，层层推进，确认要探索的目标，获得答案。

2.5.2 系统存在的问题

①：但数据中关键性文字信息在视图无法直接获得，需要结合后台数据的定向查询进行分析。

②：视图中缺少对于群体与群体之间特定属性的比较功能，这样用户无法有效的得知群体之间的相似与不同。

③：视图 5 作为时间轴模块，起到总览和视图 2，视图 3，视图 6 时间参数的输入。但是当前其所能传递的信息只有员工的上班时间，我们在整个系统实现以后感觉并不够。

④：视图 4 不能同时反映多个协议（ftp,ssh 等）的信息。

2.5.2 进一步开展研究的见解与建议

①：之后可以将文字信息展现到前端再进行进一步的分析。

②：可以考虑添加比较群体属性功能的模块，或增强现有模块（视图 3，视图 7）的功能。

③：对于视图 5 可以考虑添加员工的下班时间，与上班时间相对应。这样可以提供更丰富的信息。同时实现起来也较为方便，只要再添加一组柱状图即可。

④：对于视图 4，可以考虑添加切换功能，即横轴表示每天，纵轴表示各种协议的流量信息，颜色深浅表示流量的大小。这样可以在一个视图中发现各个流量的大小情况。

3. 参考文献

[1] ChinaVis 2018. http://chinavis.org/2018/