ChinaVis 2018 挑战 1 的可视化分析方案与设计

张浩城，强志文，曹以想，熊俊，董笑菊（上海交通大学 上海 200240）

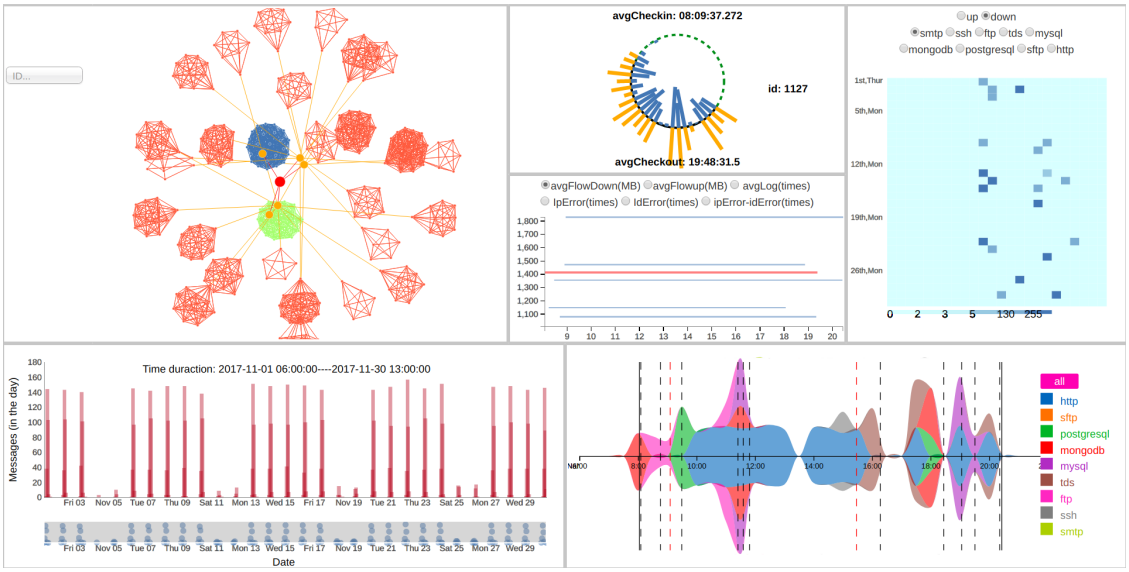

图 1　系统界面

摘要—本文针对 2018ChinaVis 挑战赛的挑战题目和数据进行深入分析，设计了一个可视分析方案，针对具体问题设计了数个视图来解析数据。题目设计对某公司员工日常行为所产生各流量协议的统计与分析，在寻找答案挖掘隐含信息的过程中采用了力导向图、矩阵、堆叠图、时空图等多种可视化分析方式。

关键字—可视化分析，交互设计，员工行为

简介

2018ChinaVis挑战赛的挑战一提供了该公司内部2017年11月共30天的多种监控数据。数据特点是数量大，维度高。我们先对整体数据进行统计分析，并针对需要探索的问题，进行数据清洗以及预处理。然后再使用可视化前端框架 d3.js 对数据进行图形表达，最后通过可视化后的系统来更深入地探索，从而得到结论。

本文的第一部分将介绍我们针对数据进行的预处理、整体界面视图、视图间的交互设计。第二部分将对当前的设计方案进行讨论。第三部分将总结本次竞赛工作的成果与不足。

1 数据处理及系统设计

1.1 数据预处理

要将公司员工分类，分析每个部门员工的日常行为并发现潜在危险的最大难点在于，无法通过30天内每天的监控数据来分析每个员工一个月内行为模式。因此，本系统通过后台处理汇总出每个员工一个月内的所有监控数据。由于已知该公司分为研发部门，财务部门，人力部门三个部门，我们还预先提取邮件中的关键词进行分析对比后将邮件基本分为三类。

1.2 界面视图及交互设计

系统如图1所示，共分为6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）、个人单日流量信息汇总模块（视图6）。

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(左下)：展示了一个月内员工出勤的时间分布。个人单日流量信息汇总模块(右下)：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。

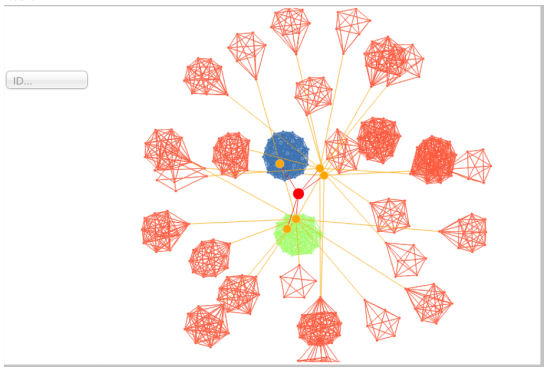

图2　员工分布模块(视图1)

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。

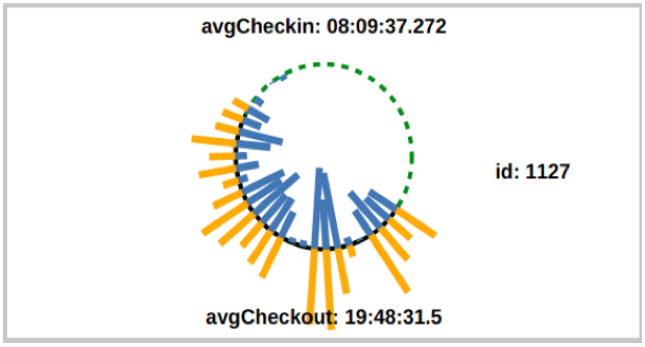

图3　个人流量展示模块（视图2）

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。

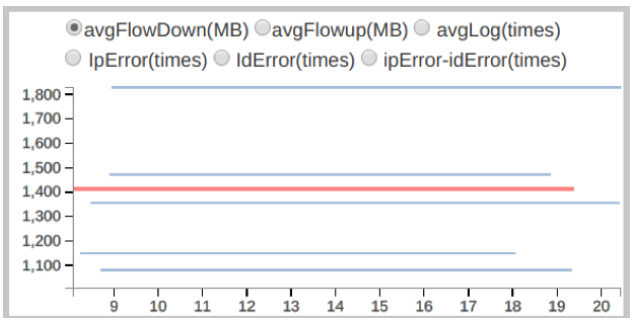

图4　群体流量分布模块（视图3）

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。

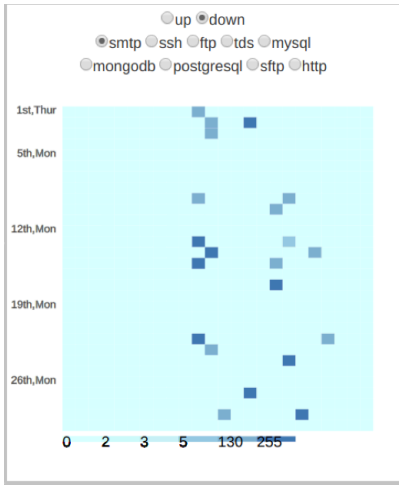

图5　个人流量时间分布模块（视图4）

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。

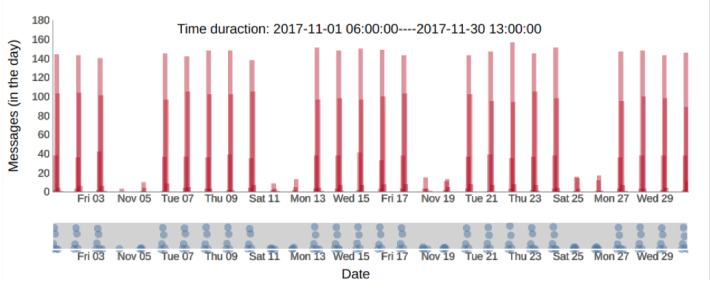

图6　时间轴轴模块（视图5）

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

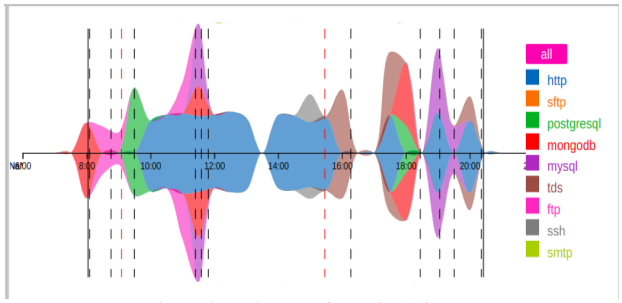

图7　个人单日流量信息汇总模块（视图6）

图7中展示的是某个id的员工某天所有流量情况的汇总(视图6)。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 讨论

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，借助不同模块的视图以及其间的联动，层层推进，确认要探索的目标，找到答案。

3 总结

视图分析内容由总览到细节，层层推进，支持以总览为起点探索细节，再从细节来验证从总览得到的信息。但数据中关键性文字信息则由在视图中精确锁定对象后结合后台数据的定向查询进行分析。之后可以将文字信息展现到前端再进行进一步的分析。

4 参考文献

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