挑战一数据可视化方案特点分析

冯晓蕾，成利，秦阳欣，李浙川，林杰（指导老师），成都理工大学

摘要—对 Hightech 互联网高科技公司一个月以来多种监控数据进行分析，构建了一套完整的可视化分析方案。通过分析公司内部邮件服务器的信息，构建关系网络图，得到公司详细的组织结构。对邮件交流和打卡记录分析得到员工日常的工作模式。再对服务器的活动、网页访问和内部网络交流情况得到公司内部更为详细的工作情况。

关键词—监控数据；可视化分析；关系网络；工作模式；

1. 概述

本次可视化方案采用从宏观到微观的方式，先宏观分析整体的工作模式和组织结构，再微观分析局部出现的异常现象。同时，可视化分析的过程，按照由独立个体到联系整体的方式逐层递进。首先对数据进行清洗处理，构建关系网络图，可以清楚地展示公司内部的部门结构和交流联系。接着对文本信息进行处理，利用词云图得到员工日常工作内容。再分别利用折线图、堆积柱形图、玫瑰图等分析员工的上下班情况、工作流程以及部门工作模式。使用甘特图、热力图、双向柱形图等分析服务器和数据库的上传下载情况、活跃度、更新情况、以及员工的登录情况。再利用饼图、瀑布图等结合 TCP 连接和网页访问数据，分析员工对外交流和数据流量传输情况。最终，分析存在的异常情况。

下面将详细介绍本次的可视化方案。

2. 部门结构可视化分析

在基于模块化 (Modularity) 算法对公司组织结构进行社区检测的条件下，使用 Louvain 算法进行公司结构划分。模块化算法公式如下，

$$Q = \frac{1}{2m}\sum_{i,j}(A_{ij} - \frac{k_i k_j}{2m})\delta(c_i, c_j)$$

Louvain 算法是对模块化算法进行迭代，每迭代一次产生一个ΔQ，直到Max(ΔQ)不在变化为止。

将公司邮箱服务器中的邮件进行清洗和筛选，除去垃圾邮件、招聘邮件、对外合作邮件和公司通知类邮件等，得到只包含公司内部员工相互交流的邮件数据。算法划分后，用关系网络图来展示各个部门之间的组织结构，如图 2-1 所示。

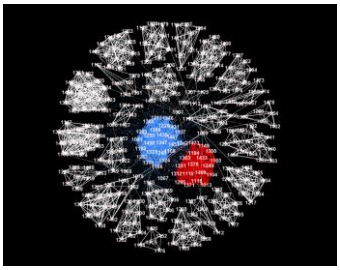

图 2-1

公司共有 299 名员工，其中研发部门人数最多。各部门内部联系紧密，高层主管只和各部门主管联系。研发部有 3 位主管，各自带领不同的研发小组，并且小组内有组长带领。财务部和人力部的人数较少，但内部交流频繁。

3. 工作模式可视化分析

3. 1 员工工作模式

得到了公司的各部门结构，对公司的工作模式进行可视化分析时，采用先整体再局部的方式。利用词云图展示不同部门的主要工作任务；折线图展示各部门工作上下班时间和数据流量传输；堆积柱形图来分析对外交流情况；玫瑰图展示网页浏览情况。可以发现相对于公司整体（9:00—18:00），财务部上下班时间偏早（8:00—17:00），而研发部由 1059 领导的部门上下班时间偏晚（10:00—19:00），如图 3-1 右侧所示。数据传输量从多到少依次为研发部、人力部和财务部。公司员工在日常工作中主要浏览公司主页、搜索引擎或技术相关网站等，如图 3-1 左侧所示。

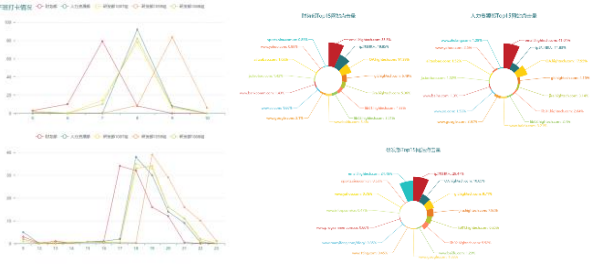

图 3-1

3. 2 公司服务器使用

对公司的服务器和数据库进行可视化分析，先整体观察服务器的活跃度可以发现，服务器每周六凌晨 2 到 3 点时分，固定更新。少数服务器在非工作日活动较为活跃，不同的服务器对应不同的应用协议，如图 3-2 所示。

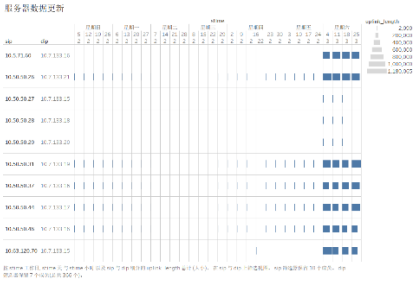

图 3-2

4. 异常事件可视化分析

对公司的整体进行宏观分析以后，得到了普遍的行为模式和清晰的部门结构。此时，对微观部分进行可视化分析。分析过程先从单独个体开始，逐步结合其他个体。可视化范围逐层递进。

4. 1 服务器异常

重点分析员工访问服务器和数据库的记录数，用堆叠柱形图展示可以发现，IP 为 10.64.105.4 的员工对端口 22 的访问量出现异常增多的情况，如图 4-1 所示。

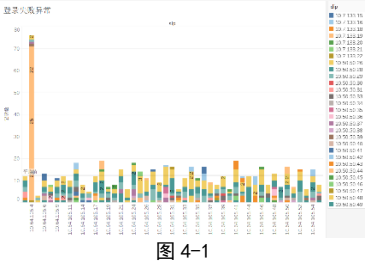

图 4-1

再对服务器和数据库的上传下载量进行分析，用双向柱形图展示可以发现，服务器 10.7.133.15 在 16 号的下载量远比其他服务器要高，如图 4-2 所示。

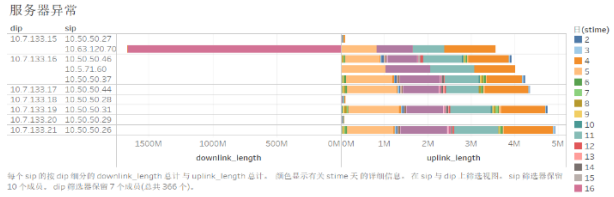

图 4-2

结合时间对服务器的活跃状态和员工访问情况进行综合分析，用甘特图展示服务器的活跃状态。在甘特图中，在 4 号非工作日这天，IP 地址为 10.50.50.44 的服务器存在大量活动异常情况，且传输协议均为 ssh，如图 4-3 所示。

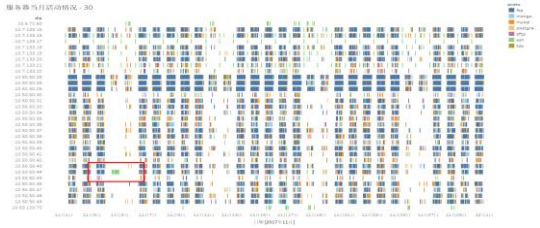

图 4-3

用热度图展示员工访问服务器情况，随着时间的变化，可以通过鼠标悬停查询员工访问服务器的记录数以及服务器活跃时的应用协议和具体时间。在热度图中，3、4、6 号这三天 IP 地址为 10.64.105.4 的员工对服务器的访问量异常增多，如下图 4-4 所示。

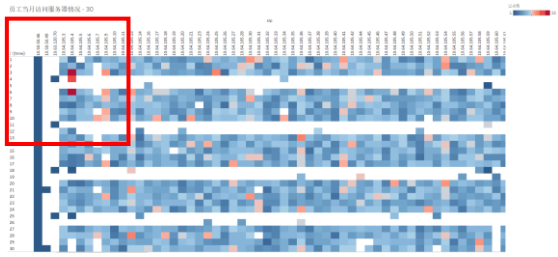

图 4-4

从个体到结合整体进行交互分析可以发现，服务器的访问异常是由于员工 1487 对 10.50.50.44 服务器的大量异常访问活动造成。

4. 2 邮件异常

对所有员工的邮件信息进行单独可视化分析。

（1）对员工接收外部邮件进行可视化分析，用树形图和饼图展示可以发现，只有研发部的部分小组组长接收大量垃圾邮件，邮件主题为广告推销、赌场注册、招聘等，其中以新葡京的广告占多数。还可以发现，员工 1487 和 1376 两位员工接收公司外部的招聘信息相对较多，如图 4-5 所示。

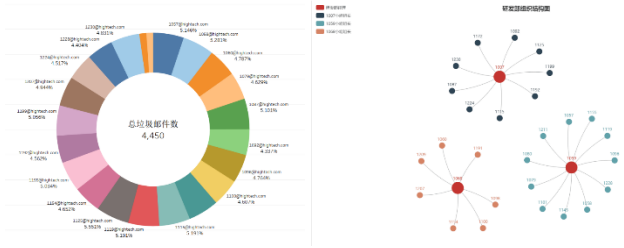

图 4-5

（2）对公司内部的邮件报警进行分析，用折线图结合时间展示不同警报内容的记录数可以发现（如图 4-6 所示）：

①在 11 月 16 号这天，公司内部名为 alert@hightech.com 的公共邮箱，多次发送了主题为“EmergencyDataBaseFatalError!“的报警。

②在 28 号这天，安全邮件崩溃次数相对于以往明显增多。

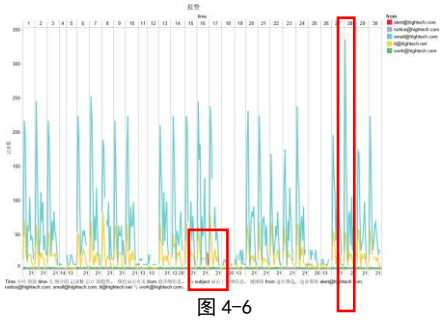

图 4-6

4. 3 员工对外部网页访问异常

对员工访问外部网站进行分析，用瀑布图展示公司员工访问的网站的比例。再使用折线图展示员工对网站访问的记录数，更直观地描述员工个体的兴趣爱好或信息需求。其中 IP 为 10.64.105.4 的员工 1487 和 IP 为 10.64.105.219 的员工 1376 访问次数远远高于平均数，如图 4-7 所示。

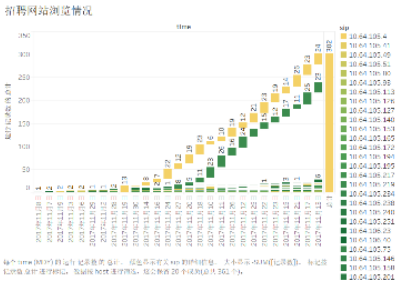

图 4-7

5. 讨论

经过完整的可视化分析，我们发现公司存在许多异常现象，联系整体分析这些异常，发现异常之间是存在关联的，一些异常之间存在着因果关系。例如，服务器出现的某些异常是由于个别员工的异常行为导致。

根据找到的异常情况，我们可以向该公司提出以下几个方案保证信息安全：

首先，是保护员工电脑文件安全、防止公司数据泄密。防止员工通过网络和移动设备进行泄密。同时，在员工离职之前，要检查并保证员工没有携带公司的情报离职。其次，还需要防止局域网共享文件服务器泄密的行为。最后，要防止垃圾邮件攻击。所谓的垃圾邮件一般具有批量发送的特征。其内容包括赚钱信息、成人广告、商业或个人网站广告、电子杂志、连环信等。

总之，企业信息安全的实现，最重要的方面是保护企业文件安全、防止公司数据泄密，这是企业信息安全的核心需求。企业管理人员除了需要高度重视，具备信息安全防护意识，还需要结合一些信息安全管理软件来保护企业机密。

6. 总结

本次可视化设计方案首先从宏观分析整体再到微观分析局部，在分析异常的过程中从个体分析再到联系整体分析，整个方案具有高效、灵活的特点，能够清晰地展示各种现象之间的联系，方便企业的管理。

参考文献

[1]《数据可视化之美》Julie Steele / Noah Iliinsky.机械工业出版社.
[2]《鲜活的数据》Nathan Yau.人民邮电出版社.