CICSC 学术大会人员分析

摘要：利用可视化技术对 CICSC 学术大会人员活动数据进行分析，得出会议的人员类型以及日程安排，并对可能出现的会议异常情况进行识别。系统主要通过停留时间长短这一关键属性将人物与地点联系起来，既包含宏观的人数变化图，也包含微观层面的人员轨迹以及关系图，可以在解读会议日程安排的同时，对相关人员活动规律进行分析。

关键词：时空分布，可视化分析，人员流动，停留时间

简介

本系统基于 CICSC 大会的人员移动数据分别进行了：单个人员特征提取、人员时空轨迹分析、宏观地点人数变化分析。所用到的可视化技术有：平行坐标系、力导向图、雷达图、折线图、条状图、散点图等。本文将重点对分析的过程、形式以及结果进行介绍。

1 数据处理方法

原始数据对人员移动情况以秒为单位进行了记录，数据量较大，不利于数据的探索与挖掘。为了得到人员在不同时间下的空间分布，以及不同地点在不同时间下的人数分布，我们对数据进行了简化：将秒数除以 60 得到分钟数，对于处于同一分钟的数据按照不同的情况统一取最后出现的数据或最初出现的数据。

完成了时间维度的整合后，我们视情况对空间维度也进行了合并，将较小的 SID 映射为更大的地点（如主会场、room1 等），这样对于探索会议的日程安排更有帮助，而且在人员时空聚类上，宏观地点显然比 SID 更有说服力。

2 可视化方案设计

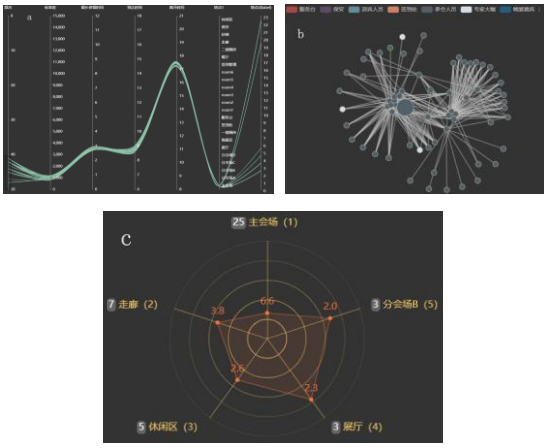
整个可视化方案主要分为三个板块：人员特征提取，用于对人员进行初步聚类，为后期数据的进一步整合提供方向；人员轨迹分析，用于对人员的具体分类以及移动规律的总结；人数时空变化分析，用于探讨会议日程安排以及会议异常情况的分析。

2.1 人员特征分析

在对人员进行聚类时，我们发觉仅仅是通过人员移动的轨迹的相似度大小来对人员进行分类并不可靠，于是我们决定对人员的时空分布数据进行进一步的整合，提取出了以下维度：频次（停留过的不同 SID 的个数，考虑到仅仅是路过并不能代表什么，我们过滤掉了停留时间小于一分钟的 SID）、标准差（针对不同 SID 的停留时间）、最长停留时间（针对宏观地点而不是 SID）、到达会场时间、离开会场时间、停留时间最长、第二长以及第三长的地点。

为了方便直观地观察各个维度的人员分布情况，我们选择了平行坐标系进行表示，在平行坐标系中可以很方便地对人员进行筛选。

我们允许对选定的人员进行更加深入的分析，以了解可能出现的异常情况。比对每个人员之间的轨迹相似度（此处表现为处于相同 SID 的分钟数），我们可以获取人物的人际关系图。将三天的数据进行合并表示，我们可以得出特定人员的 Top5 地点及其对应的停留时间。



(a)平行坐标系(b)关系图(c)雷达图

图 1 人员特征提取板块

2.2 人员轨迹分析

直接对人员轨迹进行观察是最先被想到的可视化思路，可以很直观地反映某一群体的移动情况，可以为后续数据的深度挖掘提供思路。在 Tableau 中我们可以对轨迹进行时间轴动态显示，可以指定观察的天数和人群。

有了最基本的视图，我们就可以在此基础之上进行更深入的探讨，我们接下来的分析主要基于两个思路：（a）针对人员分析地点的出现频率、（b）针对地点分析人员的停留时间，最终都是为人员的聚类以及总结人员移动规律服务。

属于思路（a）的视图：人群每天的 Top1 地点分布（图 a）、三天总共的不同 Top 层次的地点分布（图 b）、人员所经过地点的时间分布（图 c）。属于思路（b）的视图：到达与离开会议现场的时间分布（图 d）、人群在指定地点的停留时间分布（图 e）。

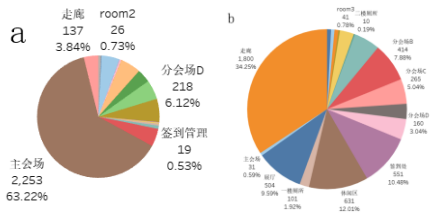


图 2 人员轨迹分析板块

2.3 人数时空变化分析（会议日程分析）

对会议日程的分析更多的是基于地点和时间查看人数的变化，比起单个人员的来去，我们更关心总体人数的涨落。最基本的视图是基于时间轴的人员流动图，如图 3。

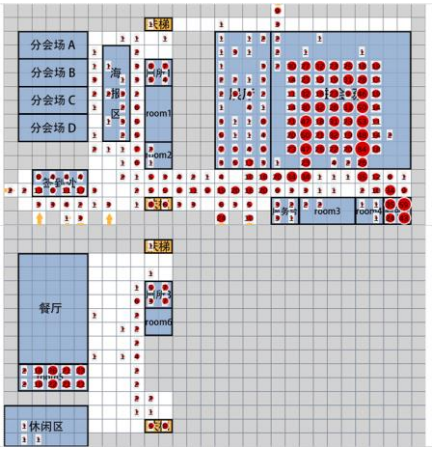


图 3 人员流动图（Day1-12:40:40）

在此基础上我们还用折线图表示指定场所的人数随时间变化情况，这也是我们区分会议日程安排的主要视图，如图 4。

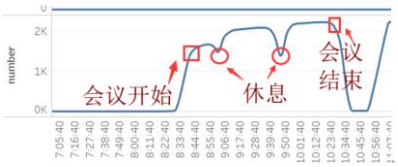


图 4 主会场人数变动

2.4 异常情况分析

针对异常情况的分析，可以从多个维度进行，首先在对人员进行分类的时候，可以发现已经确定为某一类人员的群体中，有时会出现与群体内的大部分人特征值（除了用于分类的关键特征）严重不符的人员，这时候往往需要我们利用关系图、人员轨迹图等视图对其数据进行深入探讨。

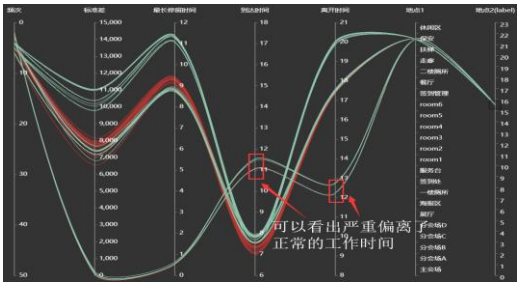


图 5 保安岗位迟到早退情况

如图 5 所示，保安岗位中存在出勤时间与其他人严重不符的人员，而且经过查看这些人中第三天与第二天迟到的人数相同且为不同的人，可能有私下换班的情况。

总体来说，异常情况一般是通过平行坐标系首先发现，例如有些数据往往过于极端，最容易引起怀疑。筛选出标准差在 0 值附近的人员，如图 6。

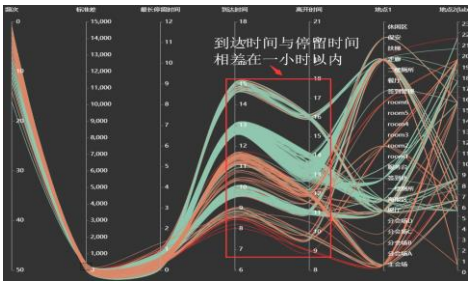


图 6 到场停留时间较短

除去缺勤的工作人员后，对以上数据中的 ID 进行查重，并未发现重复值，说明他们仅有一天出现了这种异常情况，由于该群体的其他维度如 Top1 地点以及到达的时间各不相同较难统一，又考虑到如此大型的会议必定需要不菲的注册费，平常人不会无故浪费，故猜测这些人应该是到达现场后又有另外的重要事情需要处理，所以才会立即离开。

3 可视化方案总结

3.1 简单-繁杂-简单

本次可视化是一个逐步递进的过程，开始时仅仅只有一堆原始数据，在此基础上经过简单的处理得到了人员轨迹图和整体的人员流动图，再从对人员活动轨迹的观察逐步抽象出可以反映人员活动特点的几个维度，也就有了平行坐标图。在对某些个体进行深入挖掘之后又衍生出了花样繁多的各类图表，再经过进一步的合并化简，得到最后的三大板块。

3.2 实用性

本可视化系统在不同的维度上对数据进行了表示，所有视图都是为了解决某个具体的问题而做，看此繁杂，实际上都具有内在的联系。并且本系统将较为常用的视图都置于板块的首页位置，方便大多数用户的直接使用，而当用户需要探索更深入的维度时，本系统的其他视图也可以满足需求。

在可读性方面，本系统的视图均采用大小、角度、颜色等直观的表示方法显示数据的大小，例如人员流动图原本设计为颜色标识的热力图，但考虑到颜色的渐变可能并不如形状大小的变化容易观察，才改为最终版中的样子。

3.3 可扩展性

本系统除了可以满足挑战一的需求以外，还提供了更多的数据维度可供探讨。例如图 7 中的针对特定地点的人员进出情况，可以用于优化场地开放的时间配置。

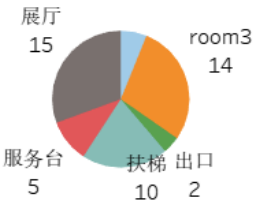


图 7 主会场 10:29:40 的人员流向