北京航空航天大學

高性能计算课程结课报告

小组成员及项目分工：

屈杨（ZF1721252）：数据搜集及数据处理
武欣（ZF1721338）：　数据搜集及可视化
王莽（ZF1721314）：数据搜集及报告撰写

目录

1 引言

1.1 时空数据的概念及特点

随着科学技术的快速发展,人类对自身生活环境的探索已经不仅仅局限于周围的世界，探索空间的外沿急剧扩展，已经遍及地球各个角落、各个圈层,并延伸到外太空。因此，如何表述人类活动的客观世界和活动特征，已经成为了科研机构和人员研究的热点和重点。伴随着计算机技术的发展,如何利用计算机模拟和表征客观世界和人类活动，无疑也为学者提供了广阔的研究空间。

伴随着人们探索空间的过程，各种信息的获取范围也从局部地面、全球地表、地球各个圈层扩展到地球内外的整个空间，从原有二维平面空间基准逐步演变到三维空间基准，进而演变到反映地理空间对象时空分布的四维空间基准。时空数据是指具有时间元素并随时间变化而变化的空间数据，是描述地球环境中地物要素信息的一种表达方式。这些时空数据涉及到各式各样的数据，如地球环境地物要素的数量、形状、纹理、空间分布特征、内在联系及规律等的数字、文本、图形和图像等,不仅具有明显的空间分布特征，而且具有数据量庞大非线性以及时变等特征。

同时具有时间和空间维度的数据，现实世界中的数据超过 80%与地理位置有关。时空大数据包括时间、空间、专题属性三维信息，具有多源、海量、更新快速的综合特点。

时空数据的特点主要体现在 5 个方面：

1）时空数据包含对象、过程、事件在空间、时间、语义等方面的关联关系。

2）时空数据具有时变、空变、动态、多维演化特点，这些基于对象、过程、事件的时空变化是可度量的，其变化过程可作为事件来描述，通过对象、过程与事件的关联映射，建立时空大数据的动态关联模型。

3）时空数据具有尺度特性，可建立时空大数据时空演化关联关系的尺度选择机制；针对不同尺度的时空大数据的时空演化特点，可实现对象、过程、事件关联关系的尺度转换与重建，进而实现时空大数据的多尺度关联分析。

4）时空数据时空变化具有多类型、多尺度、多维、动态关联特点，对关联约束可进行面向任务的分类分级，建立面向任务的关联约束选择、重构与更新机制，根据关联约束之间的相关性，可建立面向任务的关联约束启发式生成方法。

5）时空数据具有时间和空间维度上的特点，实时地抽取阶段行为特征，以及参考时空关联约束建立态势模型，实时地觉察，理解和预测导致某特定阶段行

为发生的态势。可针对时空大数据事件理解与预测问题，研究空间大数据事件行为的本体建模和规则库构建，为异常事件的模式挖掘和主动预警提供知识保障，可针对相似的行为特征，时空约束和事件级别来挖掘事件模式并构建大尺度事件及其应对方案的规则库。

1.2 报告的主要研究内容

本文根据时空数据的便捷性，应用相关技术对中国科学院“百人计划”的人才进行研究，从他们的性别、年龄、籍贯，研究所分布等方面进行着手分析挖掘，会得到一些有趣的结果，分析这些人才的分布是否与性别年龄，籍贯等一些看似无关的因素有关联。

2 项目完成情况

2.1 数据的收集

本文主要针对中国科学院“百人计划”的人才进行研究分析，人才包括所有曾经得到称号的人，即从第一次颁发“百人计划”起，到 2018 年为止。人才的数量很多，需要查询的相关属性也比较多，使用人工查询需要大量的人力物力财力，而本组人数仅三人，所以采用现今流行的爬虫方法，将这些人才的各方面信息挖掘出来，储存在简易的 excel 表格中。

在网页上，有的人才信息不全，或者书写格式不统一，所以经爬虫搜集的信息或多或少会缺失，针对这些问题，我们只能采取人工搜索的方法将信息补全，但也有一些人才的信息属于机密，所以会查不到，对于这种现象，我们也无能为力，但所幸这样的人才属于少数，对研究的结果不会产生太大的影响。

2.2 特征的构建

本文是对时空数据进行可视化分析，所以本文所搜集信息的属性主要与时间和空间相关，包括人才的性别，年龄，出生年份，籍贯以及研究所的分布。

2.3 数据的分析及结果

本文数据主要在运用高性能计算课堂上老师所传授的几种技巧和方法进行数据分析，然后将可视化结果输出在网页上，具体代码见文件夹。

2.3.1 从性别上进行分析

用饼图记录性别的分布，这样更能直观的看出男女的比列。男女比列的饼图如图 1 所示，红色区域为女性所占比例，深蓝色区域为男性所占比重，从图中不难看出，男性比例大约占 90%，女性为 10%左右，说明了男性成为中国科学院“百人计划”人才的可能性极大。所以如果想成为中国科学院“百人计划”人才，男性还是占据绝对的优势。

这可不是危言耸听，能出现这种结果，肯定有很大的依据，不妨猜测一下，

在四五十年前，中国重男轻女情节比较严重，导致男性受教育的比例会大大增加，而女性受高等教育的就相对很少。即使受过教育，多数男性事业心较重，能继续把科研进行到底，而多数女性则会成家养儿女，放弃科研，这也会使女性的比例大大降低。在智力上来说，男性的智力普遍比女性的要好一点。

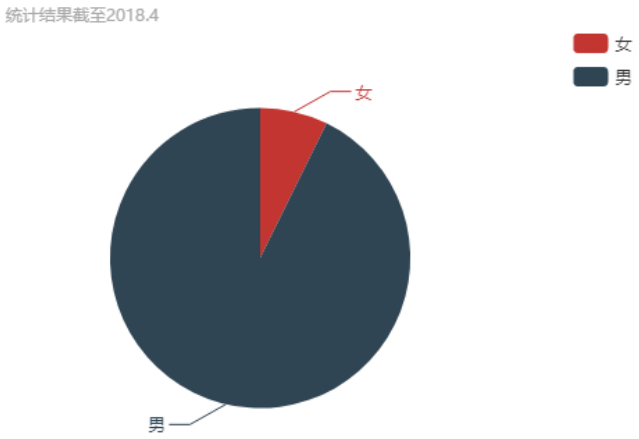

图 1 性别统计图

2.3.2 从年龄上进行分析

用饼图记录年龄的分布，很直观的看出各个年龄段所占的比重。如下图 2 所示，红色区域为小于 40 岁的，深蓝色区域为 40-50 岁之间的，浅蓝色的 区域为 50-60 岁之间的，黄色区域为大于 60 岁的。从图中不然看出，50-60 岁的占最多，40-50 岁的占第二，其他的就相对较少。经分析可知，40-60 岁的人成为中国科学院“百人计划”人才的可能性最大，换句话说，只要熬到 40-60 岁之间，就有很大希望成为中国科学院“百人计划”人才，所以正在奋斗的年轻人不要放弃科研，等年龄到了就会有很大机会。更细一点，如图 3 所示，显示出更具体的年龄分段，其体现规律和图 2 一样。

得出这条规律也是要有依据的，在 40 岁之前，人需要不断地积累经验，广泛的学习，等到 40 岁朝上的时候就很有可能一朝闻名。

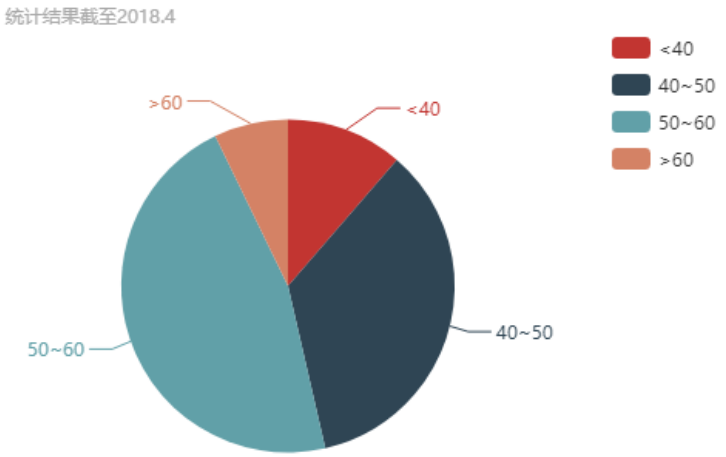

图 2 年龄统计图

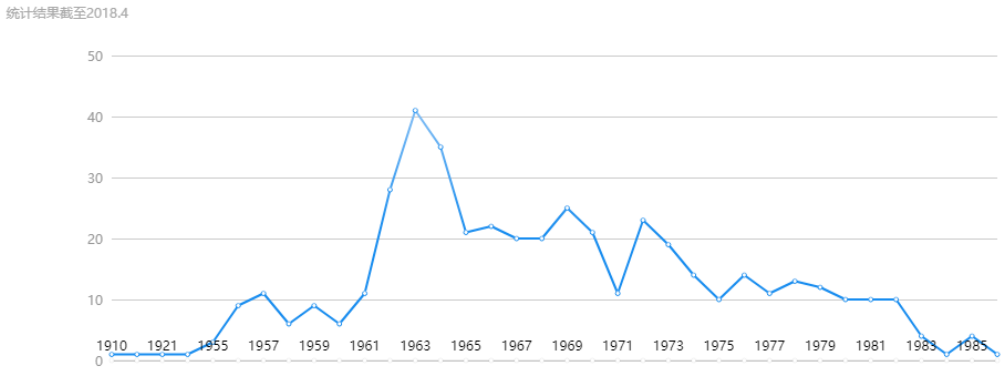

图 3 出生年份统计图

2.3.3 从籍贯上进行分析

用热力图记录籍贯的分布，可以很直观的看出人才分布。如图 4 所示，按照南北划分，南部的人才居多；按照东西划分，东部的人数居多；按照省份的划分，江苏省，浙江省，江西省，湖南省人才分布相对较集中。所以可以总结下，从上述这些地方出中国科学院“百人计划”人才的可能性最大。

导致这样的原因有很多，其中财富的划分相对较为明显，在人才比较集中的区域相对来说很富裕，生活质量会高很多，叫也相对更好一点，导致这些地区出中国科学院“百人计划”人才的可能性最大。

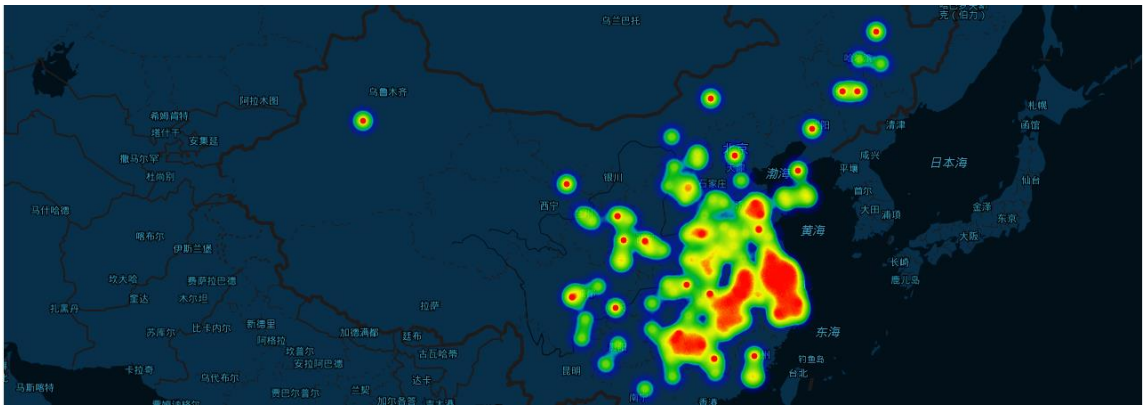

图 4 籍贯热力分布图

2.3.4 从研究所分布上进行分析

用柱状图来记录研究所的分布情况，从图中可直观看出人才所在研究所的分布情况。如下图 5 所示，很明显可以已看出，物理研究所产生中国科学院“百人计划”人才的可能性最大，其次是寒区旱区环境与工程研究所、工程过程研究所，再次是数学与系统科学研究院、上海药物研究所、生态环境研究中心。可以总结下，从事物理研究的人最有希望成为中国科学院“百人计划”人才。话句话来说，要想成为中国科学院“百人计划”人才，物理研究所是一个很大的突破口。

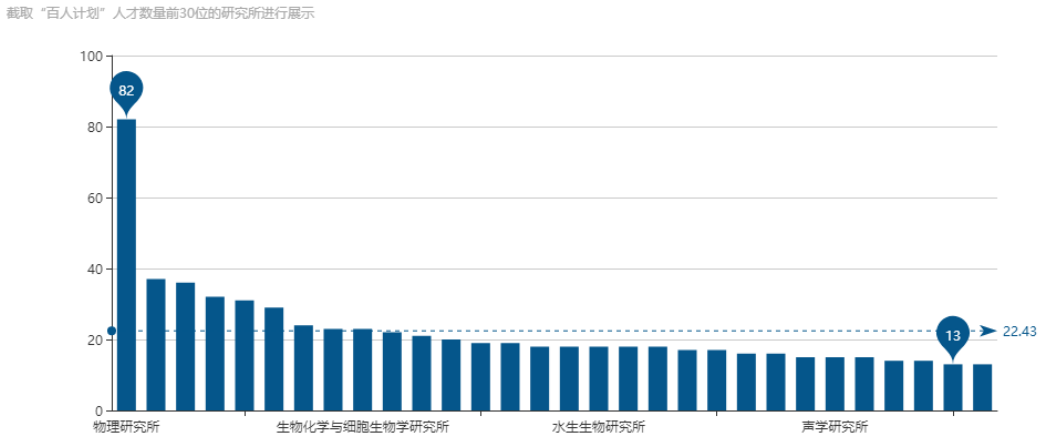

图 5 研究所分布图

3 总结与展望

本文应用高性能计算的各种方法，从性别、年龄、籍贯以及研究所分布四个方面对中国科学院“百人计划”人才分布进行可视化研究，发现其分布规律。成为中国科学院“百人计划”人才的可能性最大规律主要有：男性，年龄在 40-60 岁之间，中国东南地区，从事物理研究所工作。

本文从时空数据出发，研究了成为中国科学院“百人计划”人才可能性的因素，但所选的属性有限，希望能从更多方面对这一研究进行下去。这样会发现很多有趣且不可思议的规律。