可视化与可视分析：项目报告

孙豪
学号:516030910362

1 介绍

该项目是针对第十九届中央委员的统计数据进行可视化和分析，展示说明这 204 名委员 (或部分) 的属性特征，或依据委员个人履历分析其调任升迁的生涯轨迹，或展示委员在职场上是否曾有共事等信息。

2 数据的描述和分析

2.1 数据概述

数据的文件包括两个，第一个文件是记录 204 个委员的各种信息：包括民族、出生地、毕业院校等基本信息，以及履历，履历是格式为时间加事件的多段记录。第二个文件是全国各省市地区经纬度数据，用作地图可视化使用。

2.2 数据的维度、规模、特性

事实上我用的数据主要指第一个文件中的数据。在第一个文件中，他至少包含了如下十个维度：姓名、性别、出生地、出生年份、毕业院校、专业背景、担任职务、履历、履历中对应的时间和事件。这些维度的数据如果我们每次都只选用两个维度来作图分析的话，不仅没有全局意识，没有一个整体把握。但我们又难以融合到一个图表中。所以我们需要考虑多个视图，还要考虑一个视图中尽可能融入多个维度。

在数据规模方面，该数据的规模比较小，在 Excel 表中 204 行对应 204 个委员，列也就是对应委员的各种属性（维度）。但注意在履历的一栏中其实有大段的文字，在处理时可能需要进行分割和提取关键字（信息抽取）。

该数据的特性有以下方面

- 维度比较高，需要我们从多方面的角度去考虑委员总体的统计数据、委员单人的各项统计数据以及部分委员之间的比较和相似点分析。
- 大多数维度的数据非数值，难以进行数值大小上的比较。也就是大多数据是 Nominal level，如专业背景和毕业院校是仅有一个名字数据的。所作的数学算法仅有集合 (set) 的操作。少部分数据是 Ordinal level，可进行数值比较，如年龄和时间。没有现成数据是直接以数值给出的。
- 格式并非完全统一。如果要进行多委员的数据统计，还需要进行大量的数据清洗。

2.3 开展的分析点

1. 对于单个委员而言：

   - 履历和对应的时间是该委员的一个重要特征，必须要展开分析。对于单个委员而言：他的整个履历是一个 Time Series Model, 但是另一个维度是事件，该数据是无法用一个有大小比较的纵坐标来的。考虑到这些事件是离散的，几乎无关联的。我们可以采取时间线的方式记录这些事件。
   - 我们注意到对于某一个特定委员而言，他的升迁轨迹是值得研究的。我们可能会考虑他的职位是怎么一步一步改变，也可能会考虑到由于工作安排该委员的工作地点会不断的改变。所以，我们可以采用地图可视化来进行委员的迁移分析。这也是对于中央委员一个特有的分析方式，因为其他行业人员难以会有这样比较频繁的省份迁移。

2. 对于多个委员而言：

   - 注意到委员们基础属性中有很多可统计和比较的属性。例如我们可以考虑委员中民族为汉族的比例，也考虑 60 后和 50 后的比例，还可以考虑这些委员们的学历怎么样。针对这个角度，我们要把所有委员们整合起来，通过计数的方式来统计这些基础属性的各类别的数目。通过此做法我们可以分析出委员们的特定类别人员占比。
   - 在上一点中我们是把所有委员都统计起来，这样每个人员都只成了一个计数。丧失了其他信息。那么我们可以考虑在各种类别中加入这些人员，进行聚类分析。例如我们可以通过聚类找到某些委员是否是校友（同一院校毕业），还可以展示出出生在某一个省份的所有委员。

3 数据可视化设计

3.1 查询设计

我认为，204 个委员的统计数据。204 这个数字说大不大，说小不小。如果我们把所有委员的信息都列在我们的可视化设计上，那么必然会使得视图非常繁杂，难以让人找到。但是我们把一个委员当作一个计数 1，仅仅展示所有委员显示出的统计数据，那么会在这个过程中损失掉许多有用数据。

所以，我认为可以通过查询的方式使得用户可以自定义想要查看数据的委员，这个查询方式可以包括最基础的输入查询。还可以是通过类别和目录查询。例如，可以查询 1964 年出生的所有委员。这也就形成了和用户的交互以及图表之间的联动。

同样的，因为类别比较多，如果每个类别都做一个图表平铺到整个网页，也会大大降低用户的使用体验，所以可以采用交互的方式让用户通过按钮点击的方式切换统计类别。

该方法的好处是可以在满足用户查看需求的情况下，更精准地显示出用户想要查看的内容。但也有两个难点，一是查询需要前后端交互。二是对用户的字符串输入需要进行异常处理。

3.2 多角度和多维度的统计视图

在我们要整合出所有委员基础属性的统计数据时，关注点转移到了多个维度和多个角度的委员综合统计上。这时用户会期望不仅是一个维度，而且可以切换维度来查看多个不同的

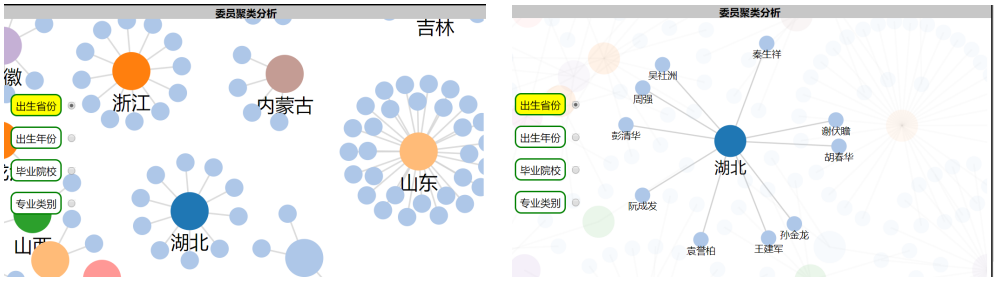
图表，用户也会期望不仅仅是一个类型的图表，例如可以看到饼状图来表示某一类别的占比，柱状图表示某一类别的具体数目。

该方法的好处是可以显示出有更多统计数据的图表，也就是可以展示出更多维的数据。难点在于维度的切换以及图表的切换。

3.3 动态的轨迹图

在设计委员的升迁轨迹时，很容易就想到在地图上作点表示委员的各工作地点。但难点在于，如何让用户了解到这是一个时间序列，也就是委员在各点出现的时间顺序先后。可能会有设计是直接在对应的地点标出年份。但这不够直观地展示出先后顺序，更不能展示出轨迹所在。这个时候就需要把时间线和地图展示融合在一起。我设计了动态的轨迹图，在地图中用一个小符号表示委员，并在地图上画出了该委员的转移工作地点的轨迹。并且，我用动图来表示整个过程，这可以让用户轻而易举地了解到先后顺序，并且对地图产生更大的兴趣。

该设计的难点有很多：第一点是数据的获取。事实上在原数据中是没有整理出履历中各年份的工作地点的。这需要我们从履历的每一条信息中提取出工作地点。因为一个一个抽取太耗时，这需要用到信息抽取的技术，提取出关键字，这个过程会比较复杂。第二是要设计一个动图，这需要一定的编程技巧来使得静态的图形看起来在动。同时由于涉及到地图和 GeoJson，这本身就是一个比较有难度的工作。



(a) 视图略微放大，未进行点击操作　　(b) 双击省份后，显示出该省份出生的所有委员名

图 1: 委员聚类视图

3.4 聚类分析

很多时候用户的关注点会在某些委员们的共同点上，例如某些委员会不会是老乡，某些委员曾经是不是校友，是否曾经同事过。那么这个时候就需要通过聚类分析。我把 204 个委员抽象成 204 个圆，这个圆点通过他们拥有的相同的类别（大圆）联系在一起。

这个设计有很多好处。第一是为查询提供了另一个接口，大多数人是不会认识全 204 个委员的，刚刚的查询部分又讲到我们不可能把委员们的名字全部列出来。那么我们就可以用这种抽象成圆的方式来提供一个查询的接口。当用户的鼠标停留在圆上的时候，可以显示出这个这个委员的名字，点击时就等同于直接输入该委员的名字。以此形成一个图表的联动。第二，我们可以在这个视图中找到拥有同样的某一个属性的所有委员。这可能是用户比较感兴趣的地方。

当然，这个设计同样也有很多难点。最大的难点是虽然我们抽象成了圆，但是这个 204 个圆也会非常占空间，如果我们每一个圆的旁边还要加上对应的标签的话，那会显得更加交错混乱。我们需要进一步设计好这个。在第二点上，我采用的方法是隐藏所有的成员标签，仅仅

显示类别标签。用户可以通过提示框的方式了解到各个圆的含义。另外，当双击某个类别时，可以高亮该类别，并直接呈现出该类别的所有成员。此外，我还提供了该视图的放大缩小和拖动功能，这可以使得用户更容易聚焦到想要查看的内容。效果如下图 fig. 5所示。

4 可视化结果

4.1 前言

我已经把整个项目上传到我的 github：https://github.com/TissueC/DataViz-project，在这上面有比较详细的各视图的用法介绍以及所有代码。此外，如果仅仅只是想要浏览该网页，可以通过 http://47.101.205.176:5000 直接打开浏览（**建议用 chrome 浏览器 F11 全屏打开**），在打开网页时有任何问题可以联系[1]我。

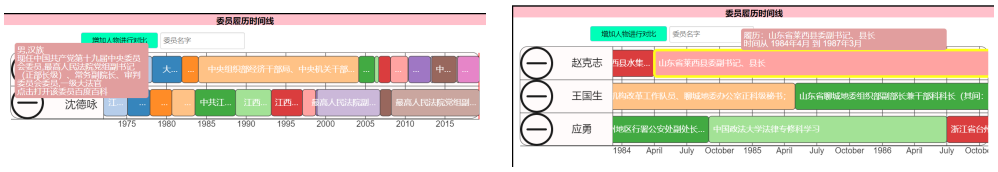
4.2 视图一：委员履历时间线

该视图是实现查询委员的主要视图。该视图完成的工作有：

- 利用时间线铺平某个特定委员的整个履历，使得用户可以查询某委员任一时间点的履历；
- 提供增删功能，方便用户对各个委员之间的时间线进行比较；
- 可提供在视图三中的地理轨迹接口；
- 提供外部百度百科接口，当用户需要委员更详细的信息时，可以点击委员名字使网页跳转到该委员的百度百科。

。

该视图显示如下图 fig. 3所示



(a) 添加委员后当鼠标悬停在名字上时，可以显示出该委员的基本信息。如果点击还可以跳转到该委员的百度百科

(b) 可以将视图放大和拖拽，鼠标悬停可以高亮，并显示出具体事件和时间

图 2: 委员履历时间线

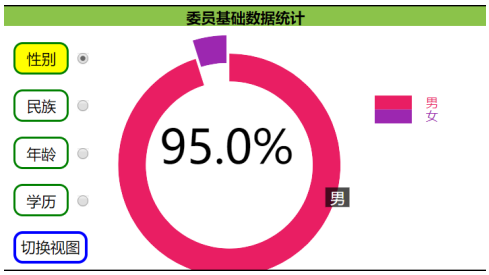
4.3 视图二：委员基础数据统计

该视图包含了 204 位委员的性别、民族、年龄、学历进行了统计。并且实现了两种图表的切换。该视图是四个视图中相对独立的一部分，它没有和其他视图产生联动（不过它自己包含了两种图表）。该视图完成的工作有：

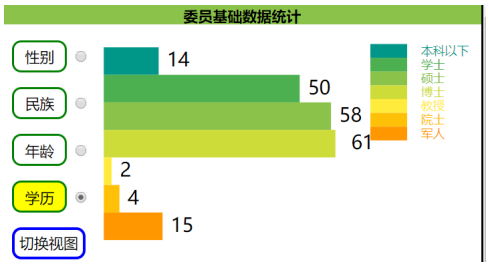
- 利用直方图展示出某一属性某一类别的具体人数；
- 利用饼图展示出某一属性某一类别人数的所占比例。

该视图显示如下图 fig. 3所示

[1]邮箱：haosun_sjtu@qq.com 电话：18217277537



(a) 选择性别属性和饼图，鼠标点击男，可以在饼图中间显示出对应的比例



(b) 选择学历属性和饼图，可以显示出所有类别的数量并形成一张横向直方图

图 3: 委员基础数据统计

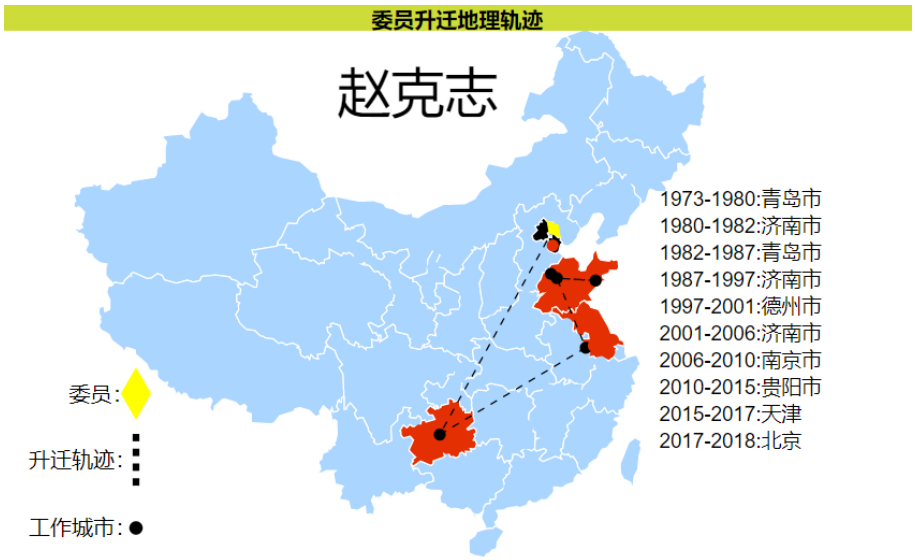


图 4: 委员升迁地理轨迹。地图动态展示出某委员的轨迹线（黄色方形标志一直在移动），高亮出该委员所经过的城市或省份。并在地图旁边以文字形式描述出经过的城市名。

4.4 视图三：委员升迁地理动态轨迹

该视图与视图一和视图四有着非常紧密的联系。该视图不能独立触发数据改变，需要经过视图一和视图四进行查询。在该部分，我把所有委员经历过的地方都整合起来并把可能会用到的城市都导入到了地图中。当鼠标悬停在地图上时，可以显示出该省份或者城市的名字。

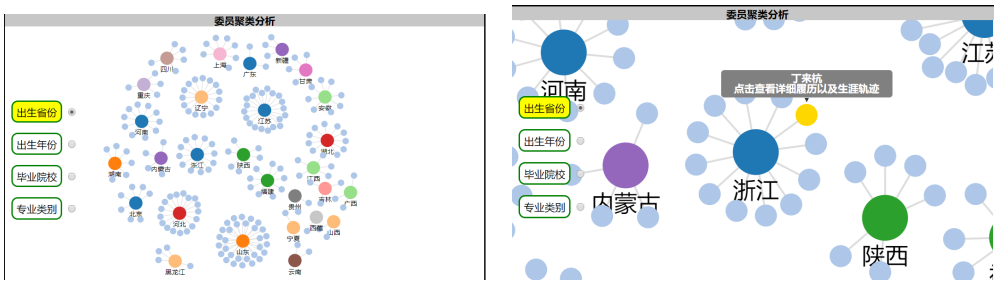
该视图完成的工作是将委员升迁的地理轨迹动态地展示出来。

4.5 视图四：委员聚类分析

该视图是利用 D3 中的力导图 (force-directed graph) 制作的。小圆圈代表的是如下图所示，该视图支持放大和缩小。在缩小的时候，所有圆点在空间中汇成一个很大的圆圈。可以调整到合适的大小以观察到所有的类别。双击类别可以高亮该类别并显示该类别所有成员；单击成员可以可以在视图一和视图三中显示出该委员的履历时间线和升迁地理轨迹。该视图完成的工作有：

- 提供一个视图一和视图三的委员查询接口，以此产生一个图表之间的联动；

- 可以根据出生省份、出生年份、毕业院校和专业类别来进行筛选查找。并可以从图中非常直观地看出各种类别的人数数量的比较；
- 可以通过缩小以概览某个属性的所有类别。



(a) 在缩小视图后。在这个视角下可以根据整个类别团的大小比较出哪个类别的成员更多。

(b) 在放大视图后，鼠标悬停在小圆上时。高亮并且显示出提示框。

图 5: 委员聚类视图

5 对于整个可视化作品的感受

在我的可视化作品中，我对某一个委员通过多个视图，进行了多个维度的描述。几乎可以保留原数据中的所有信息。整个网页我做了非常久，最后的作品虽然在很多专业人士上看起来比较幼稚，但自认为比较满意了。

整个可视化我认为最难的一点是在数据清洗上。给我们的数据格式实际上是非常糟糕的。破折号和减号混用，有些日期之后有逗号，有些没有；英文逗号和中文逗号混用；以及中英文空格的无端添加；最令人头疼的还是同一（相似）事物的说法不一造成的分类错误（例如北京市和北京，研究生和硕士，经济学、管理学和经济管理学）。我利用了我平时比较熟悉的 Python pandas 库进行了数据清洗。实际上，我还人工写了关键字白名单，最后的信息提取也是通过这个关键字白名单来提取的。

事实上数据难以清洗还是给我造成了不小的影响。因为无法提取出更多的信息，所以在视图二我只有 4 个类别，在视图四中也只有 4 个类别。我希望能够提取出更多的类别，但确实在目前所给数据的条件下，再提取出一个统一格式的类别会很耗时。

在我的可视化作品中，我加入了非常多的交互，这也是我自认为做的比较好的地方。当用户停留在几乎任意一个视图的位置，都可以获取到对应的提示框或者提示帮助。并且为了提升用户的使用体验，我加入了一些高亮、放缩和拖动，并且对用户的手动输入有错误检查与提示。我发现在这样的高交互的情况下，用户使用可视化的自由度高，体验也会更好。

我使用了前后端交互 (Python Flask) 的方式来进行数据处理和显示。虽然事实上，可以把整个需要用到的数据先转换成 Json 文件，之后完全脱离 Python 运行，甚至可以不用挂载服务器直接打开 html 文件运行。但是考虑到最后我会把整个作品放到我的服务器上，我也希望能够用前后端分离的方式做出一个更有结构性的作品，最后采取了前后端交互的方式。

最后感谢助教和老师在这次项目中的帮助，谢谢！